

Федеральное агентство по образованию  
Казанский государственный финансово-экономический институт

Кафедра статистики и эконометрики

**ПРОГРАММНЫЕ СРЕДСТВА  
СТАТИСТИЧЕСКОГО АНАЛИЗА**

Учебное пособие для студентов, обучающихся по направлениям  
521500 «Менеджмент» и 521600 «Экономика»

Казань 2005

Утверждено на заседании кафедры статистики и эконометрики  
22.02.05 г., протокол № 7.

Автор: доц. Кадочникова Е. И.

Рецензенты: доц. Кундакчян Р. М., доц. Костина Л. В.

Учебное пособие «Программные средства статистического анализа» посвящено обучению навыкам статистического анализа экономической информации на персональном компьютере. Пособие ориентировано на студентов, знакомых с общей теорией статистики и призвано помочь применять методы статистического анализа в экономических исследованиях, при решении расчетно-аналитических заданий, выполнении курсовых и выпускных квалификационных работ. Пособие состоит из двух частей. В первой части изложена технология работы с программной надстройкой «Пакет анализа» и встроенными статистическими функциями в Microsoft Excel, во второй части - возможности системы Statistica (версия 5.5) в разделах описательной статистики и анализа взаимосвязей. В соответствии с рабочей программой дисциплин «Статистика» и «Эконометрика» в пособии рассмотрены функции для вычисления средних величин, показателей вариации, корреляции, динамики; возможности построения статистических графиков, проведения дисперсионного анализа, регрессионного анализа и анализа временных рядов.

Содержание	стр.
Часть 1. Статистический анализ данных в MS EXCEL	4
1. Описательная статистика	
1.1. Стандартные статистические функции	4
1.2. Надстройка «Пакет анализа»	8
2. Дисперсионный анализ	
2.1. Стандартные статистические функции	15
2.2. Надстройка «Пакет анализа»	16
3. Статистические методы изучения взаимосвязей	
3.1. Стандартные статистические функции	21
3.2. Надстройка «Пакет анализа»	25
4. Методы анализа временных рядов	
4.1. Стандартные статистические функции	30
4.2. Надстройка «Пакет анализа»	31
Часть 2. Статистический анализ данных в системе STATISTICA	35
1. Организация хранения и обработки данных	36
2. Первичный анализ данных	40
3. Графические возможности системы STATISTICA	44
4. Регрессионный анализ	48
5. Непараметрическая статистика	55
6. Анализ временных рядов и прогнозирование	58
Список литературы	65

## Часть 1. Статистический анализ данных в MS Excel

Для статистической обработки информации в MS Excel имеется библиотека из 78 статистических функций и программная надстройка «Пакет анализа». Ниже будут рассмотрены их возможности отдельно по разделам теории статистики: описательная статистика, дисперсионный анализ, статистические методы изучения взаимосвязи, временные ряды.

### 1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

#### 1.1.Стандартные статистические функции

Работать со статистическими функциями MS Excel удобнее с помощью Мастера функций (меню **Вставка/Функция**/категория Статистические). В нем имеются следующие функции описательной статистики: СРЗНАЧ, СРГАРМ, СРГЕОМ, МЕДИАНА, МОДА, КВАРТИЛЬ, ПЕРСЕНТИЛЬ, СТАНДОТКЛОН, ДИСП, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОНА, СТАНДОТКЛОНП, ЭКСЦЕСС, СКОС, МИН, МИНА, МАКС, МАКСА, НАИБОЛЬШИЙ, НАИМЕНЬШИЙ ( см. рис. 1.1.1.).

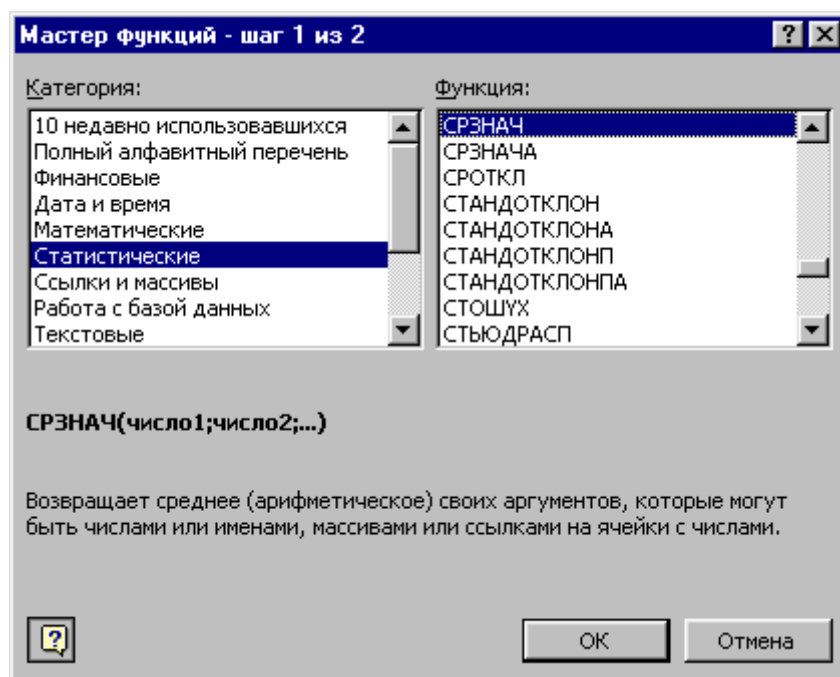


Рисунок 1.1.1. Диалоговое окно мастера функций

Остановимся на некоторых из них. **Функция СРЗНАЧ** рассчитывает значение невзвешенной средней арифметической.

Синтаксис: СРЗНАЧ (число1; число 2...).

Рассмотрим использование функции **СРЗНАЧ** для расчета численности экономически активного населения РФ в среднем за месяц (см. рис. 1.1.2).

Численность экономически активного населения в РФ, млн. чел.

	Янв.	Фев.	Март	Апр.	Май	Июнь	Июль	Авг.	Сент.	Окт.	Ноя.	Дек.
2002	71	71	71,2	71,3	71,5	71,9	72,3	72,7	72,4	72,1	71,9	71,5
2003	71,1	70,7	70,9	71,1	71,3	71,7	72	72,3				

Источник: [www.gks.ru](http://www.gks.ru)

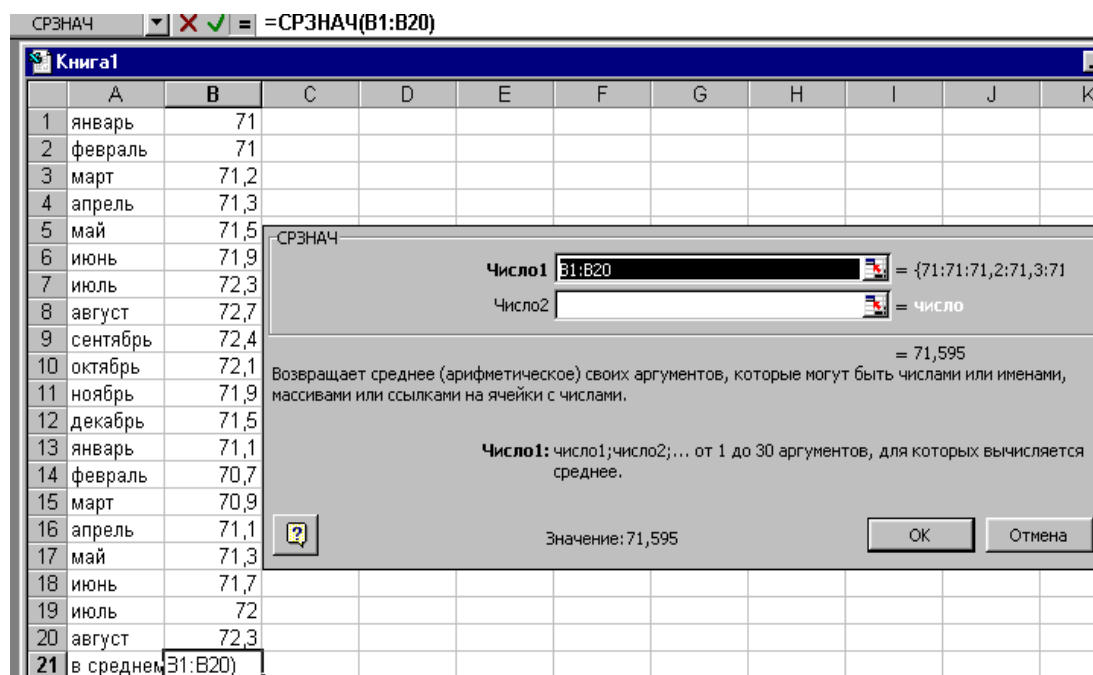


Рисунок 1.1.2. Диалоговое окно функции СРЗНАЧ

Ячейка B21 содержит формулу  $=\text{СРЗНАЧ}(B1:B20)$ , по которой рассчитывается средняя за месяц численность экономически активного населения.

Функции для расчета средней арифметической взвешенной в MS Excel нет, поэтому результат функции СУММПРОИЗВ делят на результат функции СУММ.

**Функция МЕДИАНА** рассчитывает медиану (серединное значение) дискретных данных, при этом ранжирование данных выполняется автоматически.

Синтаксис: МЕДИАНА (число1; число2;...)

В18      =МЕДИАНА(В4:В17)

	А	В	С	Д	Е
1	Численность постоянного населения в				
2	Приволжском федеральном округе, тыс. чел.				
3					
4	Республика Башкортостан	4102,9			
5	Республика Марий Эл	728			
6	Республика Мордовия	888,7			
7	Республика Татарстан	3779,8			
8	Удмуртская Республика	1570,5			
9	Чувашская Республика	1313,9			
10	Кировская область	1503,6			
11	Нижегородская область	3524			
12	Оренбургская область	2177,5			
13	Пензенская область	1453,4			
14	Пермская область	2824,4			
15	Самарская область	3239,8			
16	Саратовская область	2669,3			
17	Ульяновская область	1382,3			
18	Медиана	1874			

Рисунок 1.1.3. Результат функции МЕДИАНА (источник:www.gks.ru).

В ячейке В18 помещена формула =МЕДИАНА(В4:В17), она определила значение 1874. ( см. рис. 1.1.3)

Расчет медианы по интервальным рядам требует проведения определенных расчетов ( см. рис. 1.1.4):

ячейка В8: формула =СУММ (В3:В7);

ячейка В9: формула =В8/2 (50% поселений);

ячейка В10: формула =ПОИСКПОЗ(В9;С3:С7;1)- в массиве С3:С7 определяется номер позиции числа, которое является наибольшим среди чисел, меньших или равных середине интервала, т. е. числа 1540,5;

ячейка В11: формула =ИНДЕКС(С3:С7;В10;1) – из массива С3:С7 извлекается число, удовлетворяющее условиям поиска, сформированным в ячейке В10;

ячейка В12: формула = ЕСЛИ(В9=В11;В10;В10+1)- рассчитывается смещение на медианный интервал;

ячейка В13: формула = ИНДЕКС(С3:С7;В12;1) – отображается частота медианного интервала;

ячейка В14: формула = ИНДЕКС(А3:А7;В12;1) – указан медианный интервал;

ячейка B15: формула =ЛЕВСИМВ(B14;1) – отражается нижняя граница медианного интервала;

ячейка B16: формула =ИНДЕКС(C3:C7;B12-1;1) – отражается накопленная частота интервала, предшествующего медианному;

ячейка B17: формула = 101+49\*((B9-B16)/B13) – рассчитывается медианная численность населения, проживающего в сельских поселениях.

B17      = 101+49\*((B9-B16)/B13)

Книга1				
	A	B	C	D
1	Группировка сельских поселений			
2	Республики Татарстан			
3	без населения	35	35	
4	до 10 человек	201	236	
5	"11-50	407	643	
6	51-100	407	1050	
7	101 и более	2031	3081	
8	Итого	3081		
9	50% поселений	1540,5		
10	смещение на $\max < N/2$	4		
11	значение $\max < N/2$	1050		
12	смещение на медианный	5		
13	Частота мед.	3081		
14	Медианный интервал	101 и более		
15	Нижняя граница	1		
16	Накопленная частота	1050		
17	Медиана населения	108,8009		
18				

Рисунок 1.1.4. Расчет медианы интервального ряда

B17      = МОДА(B3:B16)

Методпк		
	A	B
1	Величина прожиточного минимума	
2	за 2 квартал 2003 года, рублей	
3	Республика Башкортостан	1775
4	Республика Марий Эл	1785
5	Республика Мордовия	1851
6	Республика Татарстан	1803
7	Удмуртская Республика	2004
8	Чувашская Республика	1760
9	Кировская область	1968
10	Нижегородская область	2041
11	Оренбургская область	1880
12	Пензенская область	1824
13	Пермская область	2163
14	Самарская область	2248
15	Саратовская область	1968
16	Ульяновская область	1857
17	Мода	1968
18		

Рисунок 1.1.5. Результат функции МОДА (источник:www.gks.ru)

**Функция МОДА** рассчитывает моду дискретных данных.

Синтаксис: МОДА (число1; число2;...)

Ячейка В17 содержит формулу МОДА (В3:В16) (см. рис. 1.1.5).

Мода интервального ряда в MS Excel автоматически не определяется.

Поэтому требуется записывать вручную ряд формул.

## 1.2. Надстройка «Пакет анализа»

С помощью данной надстройки (меню *Сервис/Анализ данных/Пакет анализа*) можно определить показатели описательной статистики, ранг и персентиль, построить гистограмму, выполнить выборку.

Режим *«Гистограмма»* позволяет построить дискретный ряд сгруппированных данных и представить их графически на диаграмме Парето. В данном режиме имеются следующие элементы управления:

1. Поле *Входной интервал* – вводится ссылка на ячейки, содержащие анализируемые данные.
2. Флажок *Метки* устанавливается в активное состояние, если первая строка (столбец) во входном диапазоне содержит заголовки.
3. Поле *Интервал карманов* (необязательное). Введите в поле ссылку на ячейки, в которых заданы границы интервалов группировки (карманов) в возрастающем порядке. Например, карману со значением 1000 будет соответствовать частота данных, меньших, чем 1000, но больших, чем предшествующий карман.. Если диапазон карманов не был введен, то набор отрезков, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.
4. Поле *Парето* (отсортированная диаграмма). Флажок позволяет представить данные в порядке убывания частоты. Поле *Интегральный процент*. Флажок включает в гистограмму график кумулятивных процентов.
5. Поле *Вывод графика*. Флажок позволяет автоматически создать встроенную диаграмму на листе, содержащем выходной диапазон.



Рассмотрим построение диаграммы Парето по данным о численности постоянного населения (см. рис. 1.2.1).

Численность постоянного населения, тыс. чел.  
(по данным переписи 2002 г.)

Субъект РФ	Численность	Интервал (карман)
Республика Башкортостан	4102,9	1000
Марий Эл	728	2000
Республика Мордовия	888,7	2500
Республика Татарстан	3779,8	3000
Удмуртская республика	1570,5	3500
Чувашская республика	1313,9	4000
Кировская область	1503,6	4500
Нижегородская область	3524	
Оренбургская область	2177,5	
Пензенская область	1453,4	
Пермская область	2824,4	
Самарская область	3239,8	
Саратовская область	2669,3	
Ульяновская область	1382,3	

Источник: [www.gks.ru](http://www.gks.ru)

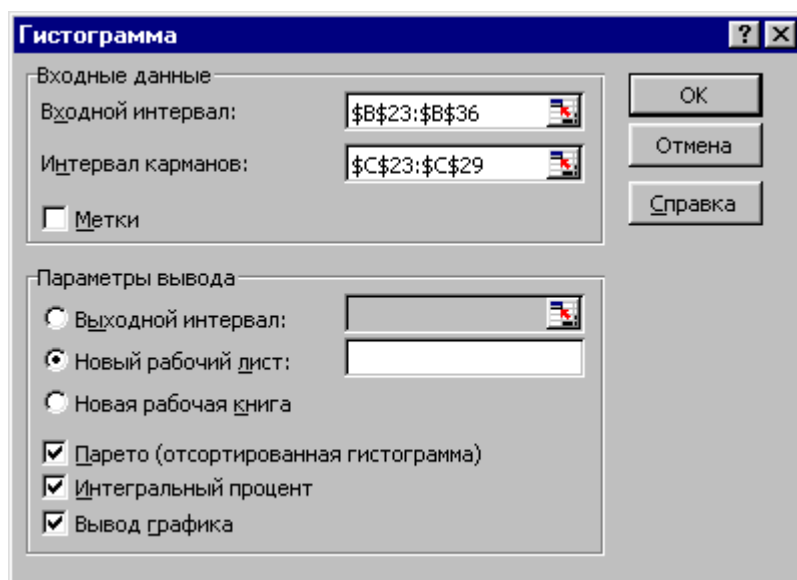


Рисунок 1.2.1. Диалоговое окно режима Гистограмма

В нижеследующей таблице представлен дискретный ряд территорий, сгруппированных по численности населения. Очевидно, что численность населения менее 1000 тыс. человек имеют 2 субъекта (Марий Эл и Республика Мордовия), в них сосредоточено 14,29% населения округа. От 1000 до 2000

тыс. человек постоянного населения имеют 5 субъектов (35,7% населения округа). Две первые группы включают в себя 50 % всего населения Приволжского административного округа.

Карман	Частота	Интегральный %	Карман	Частота	Интегральный %
1000	2	14,29%	2000	5	35,71%
2000	5	50,00%	1000	2	50,00%
2500	1	57,14%	3000	2	64,29%
3000	2	71,43%	4000	2	78,57%
3500	1	78,57%	2500	1	85,71%
4000	2	92,86%	3500	1	92,86%
4500	1	100,00%	4500	1	100,00%
Еще	0	100,00%	Еще	0	100,00%

Высота столбика на гистограмме показывает частоту каждой группы субъектов. В Приволжском административном округе преобладают субъекты с населением от 1000 до 2000 тыс. человек. Линия отражает кумулятивный процент численности населения. Например, в трех первых группах субъектов проживает 57,1 % населения округа.

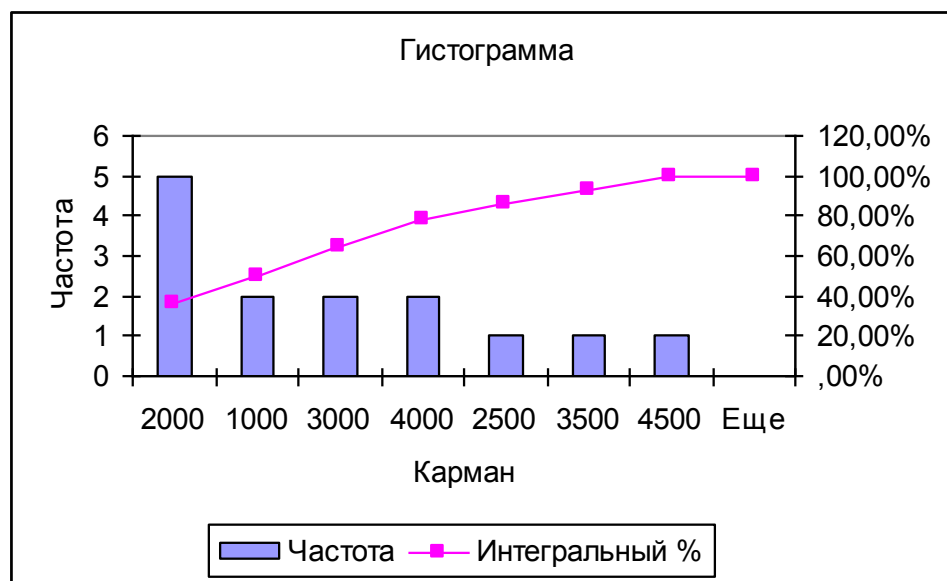


Рисунок 1.2.2. Диаграмма Парето для субъектов РФ по численности постоянного населения.

Режим «**Выборка**» позволяет формировать выборку из генеральной совокупности на основе схемы повторного собственно-случайного отбора, а также из периодических данных. Диалоговое окно режима «Выборка» содержит элемент управления «Метод выборки» (см. рис. 1.2.3).

В положении «Периодический» в поле «Период» указывают размер периодического интервала выборки. В положении «Случайный» в поле «Число выборок» указывают объем выборки.

Например, студклуб проводил среди студентов лотерею по распространению билетов на вечер КВН. Список 200 студентов размещен на рабочем листе Excel. Необходимо отобрать 10 студентов.

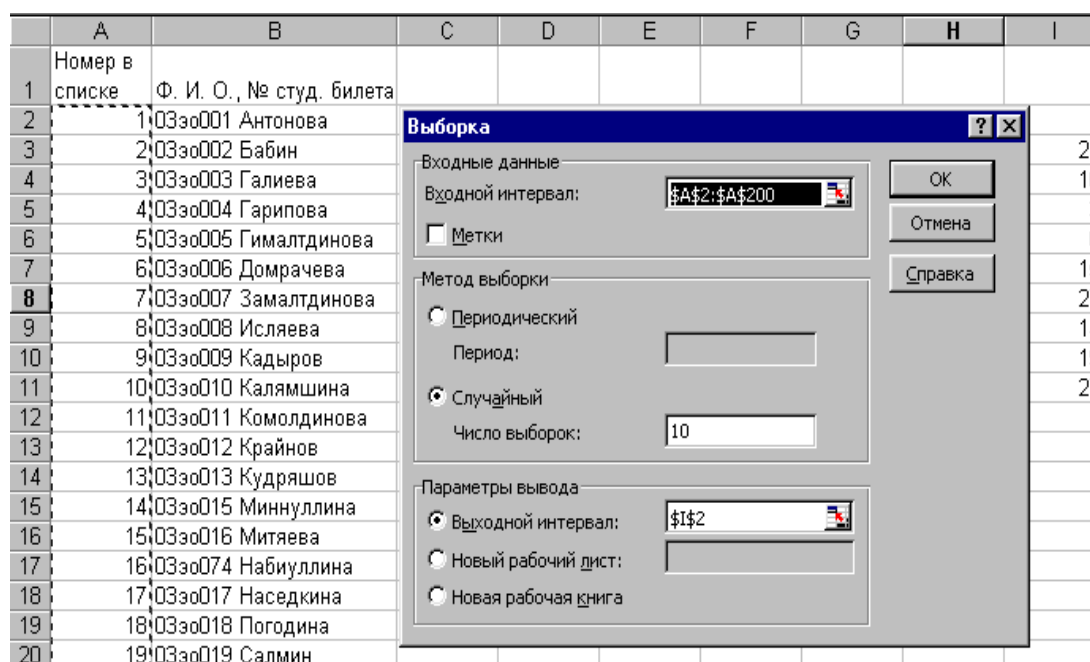


Рисунок 1.2.3. Диалоговое окно режима «Выборка».

Номера, которые оказались выигрышными, указаны в столбце I, студент с номером 27 выиграл 2 билета.

В режиме «*Описательная статистика*» выполняется расчет основных показателей положения (средняя арифметическая выборки, средняя ошибка выборки, медиана, мода, размах вариации, минимальный и максимальный элементы выборки), разброса (оценка среднего квадратического отклонения и дисперсии по выборке) и асимметрии (оценка эксцесса и коэффициента асимметрии) по выборочной совокупности. Для вывода указанных показателей надо активизировать элемент управления «Итоговая статистика». Элемент управления «Уровень надежности», установленный в активное состояние, позволяет рассчитать предельную ошибку выборки для требуемой доверительной вероятности (обычно 95%). Элемент управления «К-й наибольший» в активном состоянии включает в выходную таблицу k-е наибольшие значения (начиная с

максимального элемента выборки). Если  $k=1$ , то строка будет содержать только максимальное значение элемента выборки. Аналогично назначение элемента управления «К-й наименьший» (см. рис. 1.2.4. ).

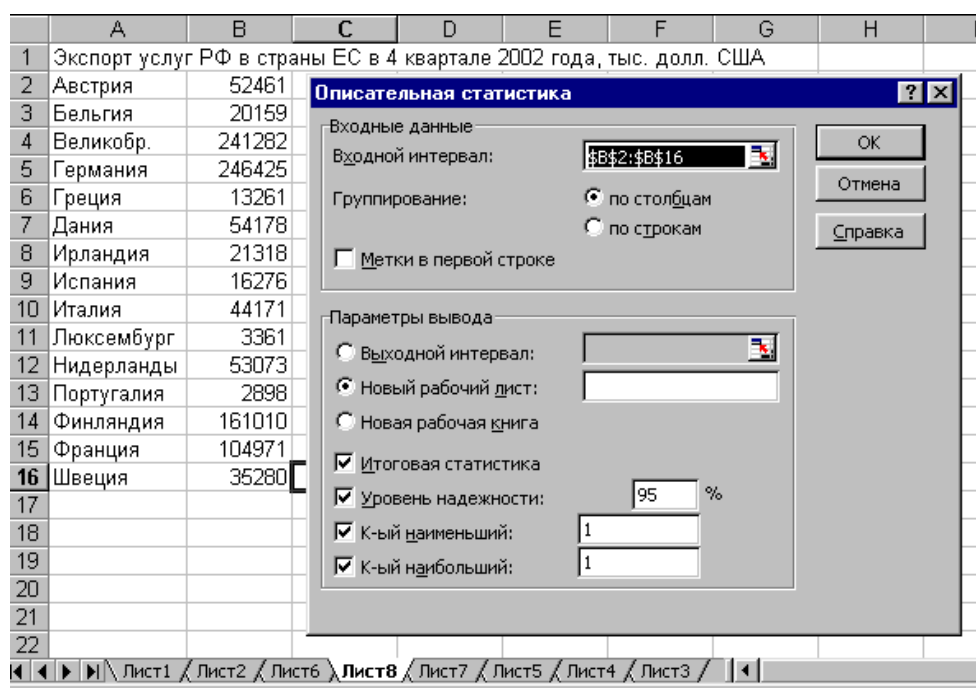


Рисунок 1.2.4. Диалоговое окно режима «Описательная статистика»

Источник: [www.cbr.ru](http://www.cbr.ru)

Основные показатели описательной статистики приведены ниже.

Столбец1	
Среднее	71342
Стандартная ошибка (средняя ошибка выборки)	20995
Медиана	44171
Мода	#Н/Д
Стандартное отклонение (среднее квадратическое отклонение)	81312
Дисперсия выборки	6611590524
Эксцесс	1
Асимметричность	1
Интервал (размах вариации)	243527
Минимум	2898
Максимум	246425
Сумма	1070124
Счет (объем выборки)	15
Наибольший(1)	246425
Наименьший(1)	2898
Уровень надежности(95,0%)	45029

На основании данной таблицы (по показателям средней арифметической выборки и предельной ошибки выборки) с уровнем надежности 95% можно предположить, что средний размер экспорта услуг РФ в страны ЕС в 4 квартале 2002 года находился в пределах от 26313 тыс. долл. (71342-45029) до 116371 тыс. долл. (71342+45029).

Коэффициент вариации, равный 114% свидетельствует о сильной колеблемости экспорта услуг в выборке. Ненадежность средней подтверждается и ее значительным отклонением от медианы выборки. Значения коэффициентов асимметрии и эксцесса, равные 1, свидетельствуют, что данное распределение имеет правостороннюю асимметрию и характеризуется скоплением членов ряда в центре распределения.

Режим «**Ранг и персентиль**» служит для генерации таблицы, содержащей порядковые и процентные ранги для каждого значения из набора данных, при этом данные упорядочиваются в порядке убывания (см. рис. 1.2.5). Ранг (R) определяет номер (порядковое место) значения случайной величины в наборе данных. Персентиль (Ti) показывает процентный ранг для каждого значения:

$$T_i = \frac{(n - R_i - (k_i - 1))}{n - 1} * 100$$

n – количество данных в наборе;

Ri\_ – ранг i-го числа, рассчитанный при условии упорядочения данных по убыванию;

k<sub>i</sub> – количество повторяющихся значений i-го числа в наборе данных.

Ранги находят практическое применение в непараметрических методах оценки взаимосвязи. Например, определим коэффициент Спирмена, используя режим «Ранг и Персентиль». Коэффициент Спирмена вычислим по формуле:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 90}{15 \cdot (15^2 - 1)} = 0,84$$

Значение коэффициента Спирмена, равное 84%, свидетельствует о сильной связи между экспортом и импортом услуг.

	A	B	C	D	E	F	G	H	I	J
1	Услуги РФ и стран ЕС в 4 квартале 2002 года, тыс. долл. США									
2			экспорт	импорт						
3	Австрия		52461	72846						
4	Бельгия		20159	32162						
5	Великобр.		241282	345111						
6	Германия		246425	398992						
7	Греция		13261	41382						
8	Дания		54178	29688						
9	Ирландия		21318	21493						
10	Испания		16276	44802						
11	Италия		44171	100719						
12	Люксембург		3361	1326						
13	Нидерланды		53073	84237						
14	Португалия		2898	3705						
15	Финляндия		161010	390285						
16	Франция		104971	141191						
17	Швеция		35280	55499						
18										

**Ранг и перцентиль** ? X

Входные данные

Входной интервал:

Группирование: ☒ по столбцам ☐ по строкам

☒ Метки в первой строке

Параметры вывода

☐ Выходной интервал:

☒ Новый рабочий лист:

☐ Новая рабочая книга

OK Отмена Справка

Рисунок 1.2.5. Диалоговое окно режима «Ранг и перцентиль».

Результаты выполнения данного режима приведены ниже.

Точка	экспорт	Ранг	Процент	Точка	импорт	Ранг	Процент
4	246425	1	100,00%	4	398992	1	100,00%
3	241282	2	92,80%	13	390285	2	92,80%
13	161010	3	85,70%	3	345111	3	85,70%
14	104971	4	78,50%	14	141191	4	78,50%
6	54178	5	71,40%	9	100719	5	71,40%
11	53073	6	64,20%	11	84237	6	64,20%
1	52461	7	57,10%	1	72846	7	57,10%
9	44171	8	50,00%	15	55499	8	50,00%
15	35280	9	42,80%	8	44802	9	42,80%
7	21318	10	35,70%	5	41382	10	35,70%
2	20159	11	28,50%	2	32162	11	28,50%
8	16276	12	21,40%	6	29688	12	21,40%
5	13261	13	14,20%	7	21493	13	14,20%
10	3361	14	7,10%	12	3705	14	7,10%
12	2898	15	,00%	10	1326	15	,00%

	Экспорт, X	Импорт, Y	ранг X	ранг Y	(Rx-Ry)^2
1	2	3	4	5	6
Австрия	52461	72846	7	7	0
Бельгия	20159	32162	11	11	0
Великобр.	241282	345111	2	3	1
Германия	246425	398992	1	1	0
Греция	13261	41382	13	10	9
Дания	54178	29688	5	12	49
Ирландия	21318	21493	10	13	9

1	2	3	4	5	6
Испания	16276	44802	12	9	9
Италия	44171	100719	8	5	9
Люксембург	3361	1326	14	15	1
Нидерланды	53073	84237	6	6	0
Португалия	2898	3705	15	14	1
Финляндия	161010	390285	3	2	1
Франция	104971	141191	4	4	0
Швеция	35280	55499	9	8	1
			Сумма		90

## 2. ДИСПЕРСИОННЫЙ АНАЛИЗ

### 2.1.Стандартные статистические функции

Рассмотрим основные стандартные статистические функции.

**Функция ДИСП** оценивает несмещенную дисперсию по выборке данных по формуле:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Синтаксис: СТАНДОТКЛОН (число1; число2;...).

**Функция СТАНДОТКЛОН** оценивает стандартное отклонение по выборке данных по формуле:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Синтаксис: ДИСП (число1; число2;...) (См. рис. 2.1.1).

**Функция КВАДРОТКЛ** рассчитывает сумму квадратов отклонений точек данных от их средней арифметической:

$$\sum (x_i - \bar{x})^2$$

Синтаксис: КВАДРОТКЛ (число 1; число 2; ...)

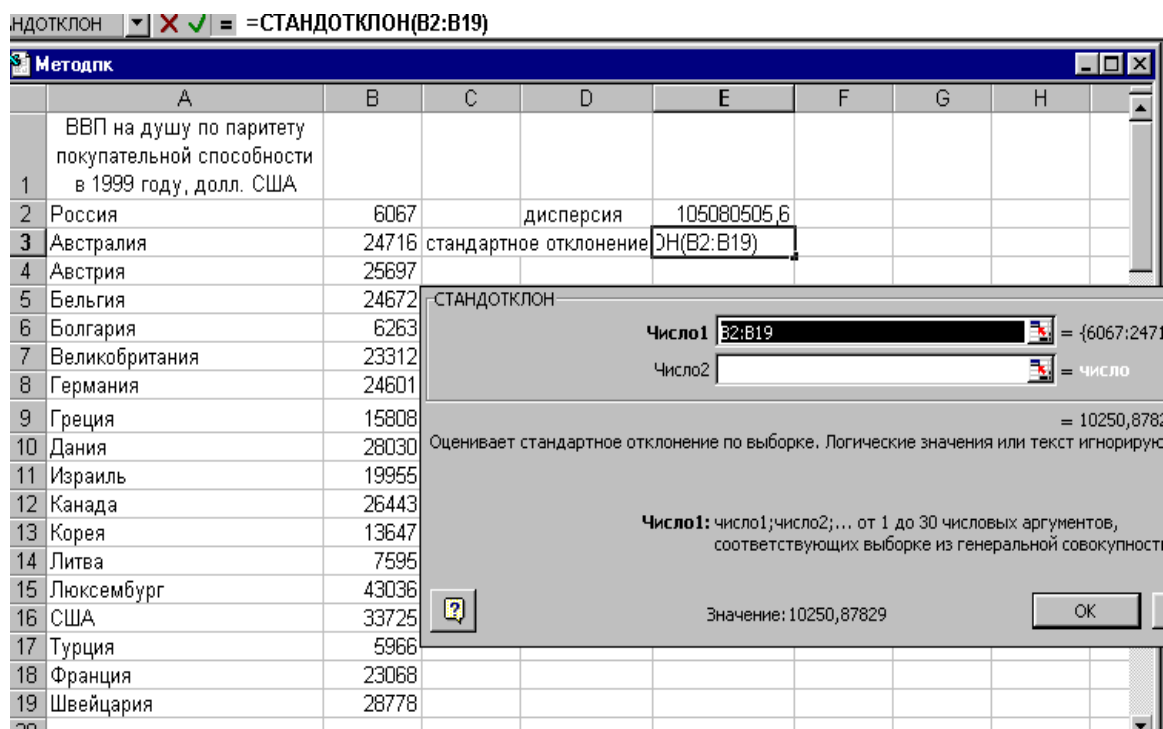


Рисунок 2.1.1. Диалоговое окно функции СТАНДОТКЛОН

## 2.2. Настройка «Пакет анализа»

В режиме *«Однофакторный дисперсионный анализ»* выполняется разложение общей выборочной дисперсии на сумму дисперсии групповых средних и средней из групповых дисперсий (см. рис. 2.2.1).

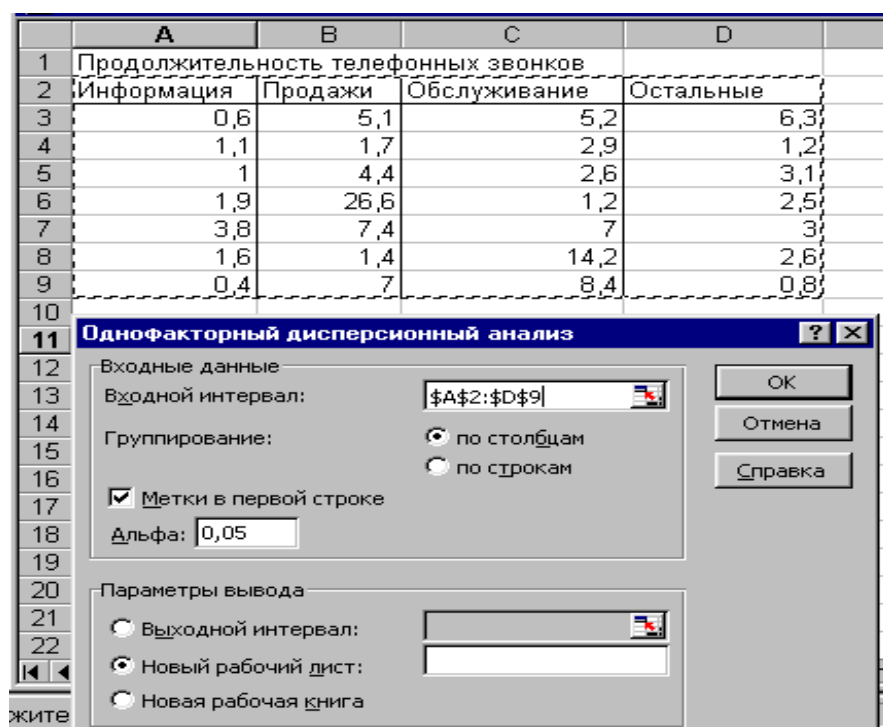


Рисунок 2.2.1. Диалоговое окно режима «Однофакторный дисперсионный анализ»



В диалоговом окне данного режима задаются следующие параметры: входной интервал; группирование; метки в первой строке; альфа – вводится уровень значимости  $\alpha$ , равный вероятности возникновения ошибки первого рода (вероятности отвергнуть нулевую гипотезу); выходной интервал.

Например, для лучшего распределения рабочего времени зафиксирована продолжительность телефонных звонков по определенным темам.

Результаты выполнения режима содержатся в таблице однофакторного дисперсионного анализа.

Однофакторный дисперсионный анализ

ИТОГИ

<i>Группы</i>	<i>Счет</i>	<i>Сумма</i>	<i>Среднее</i>	<i>Дисперсия</i>
Информация	7	10,4	1,48	1,31
Продажи	7	53,6	7,65	75,18
Обслуживание	7	41,5	5,92	19,80
Остальные	7	19,5	2,78	3,17

Дисперсионный анализ

<i>Источник вариации</i>	<i>SS</i>	<i>Df</i>	<i>MS</i>	<i>F</i>	<i>P-Значение</i>	<i>F критическое</i>
Между группами	168,19	3	56,06	2,25	0,10	3,008786109
Внутри групп	596,88	24	24,87			
Итого	765,08	27				

Из таблицы очевидно, что наибольшей средней продолжительностью (7,65 минут) и вариацией времени разговора (коэффициент вариации, равный 113%) обладают звонки по продажам.

Показатель *SS между группами* содержит взвешенную сумму квадратов отклонений групповых средних от общей выборочной средней. Показатель *SS внутри групп* содержит остаточную сумму квадратов отклонений наблюдаемых значений уровня от своей выборочной средней. Показатель *SS итого* содержит общую сумму квадратов отклонений наблюдаемых значений от общей выборочной средней. Показатель *MS между группами* содержит оценку межгруппо-

вой (факторной) дисперсии. Показатель *MS внутри групп* содержит оценку внутригрупповой (остаточной) дисперсии.

$$R^2 = \frac{168,19}{765,08} = 0,2198$$

Выборочный коэффициент детерминации ( $R^2$ ) показывает, что 22% общей выборочной вариации времени разговора связано с конкретной тематикой. Наблюдаемое значение F-статистики меньше, чем критическое, подтверждает, что средняя продолжительность разговоров незначимо различается в зависимости от тематики.

### Режим “Двухфакторный дисперсионный анализ без повторений”

Основой проведения двухфакторного дисперсионного анализа служит комбинационная группировка по двум факторам. Общая выборочная дисперсия определяется как сумма межгрупповых дисперсий по каждому из факторов (*MS строки*., *MS столбцы*) и остаточной дисперсии (*MS погрешность*). При выполнении двухфакторного дисперсионного анализа без повторений каждому уровню факторов соответствует только одна выборка данных.

Например, требуется при уровне значимости 0,05 выяснить, влияют ли на оценку качества продукции рабочие смены и поставщики исходных материалов (см. рис. 2.2.2).

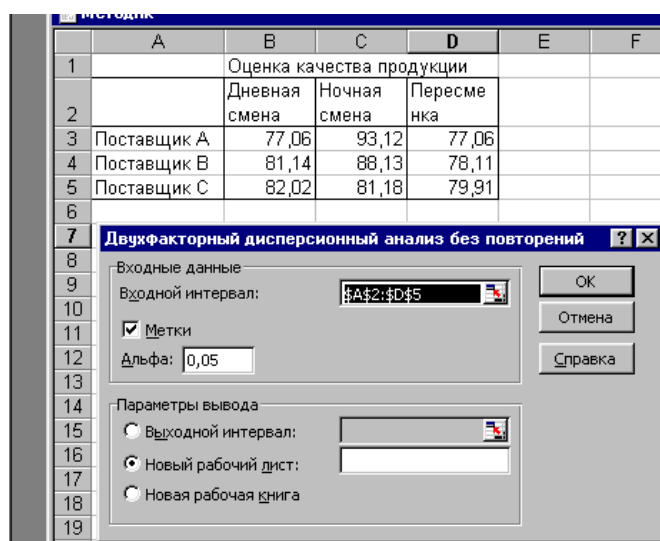


Рисунок 2.2.2. Диалоговое окно режима “Двухфакторный дисперсионный анализ без повторений”.

## Двухфакторный дисперсионный анализ без повторений

ИТОГИ	Счет	Сумма	Среднее	Дисперсия
Поставщик А	3	247,24	82,413333	85,974533
Поставщик В	3	247,38	82,46	26,4069
Поставщик С	3	243,11	81,036667	1,1284333
Дневная смена	3	240,22	80,073333	7,0037333
Ночная смена	3	262,43	87,476667	35,961033
Пересменка	3	235,08	78,36	2,0775

## Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Строки	3,923267	2	1,9616333	0,091068	0,91479479	6,944276265
Столбцы	140,8585	2	70,429233	3,2696471	0,14404458	6,944276265
Погрешность	86,16127	4	21,540317			
Итого	230,943	8				

Наблюдаемое значение F-статистики для каждого из факторов (*строки* – поставщики, *столбцы* – рабочая смена) меньше критического значения. Значит, данная выборка свидетельствует о том, что не рабочая смена, не тип поставщика не оказывают влияния на качество продукции.

## Режим “Двухфакторный дисперсионный анализ с повторениями”

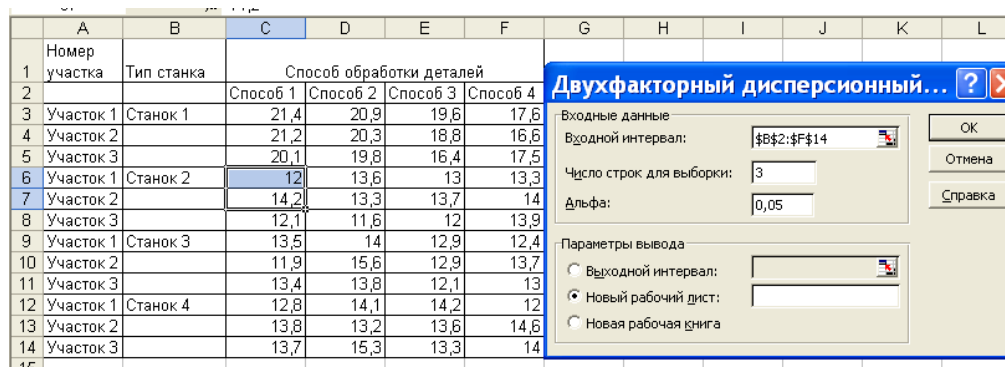


Рисунок 2.2.3. Диалоговое окно режима “Двухфакторный дисперсионный анализ с повторениями”.

При выполнении двухфакторного дисперсионного анализа с повторениями каждому уровню факторов соответствует несколько выборок данных.

Например, имеются следующие выборочные данные о выработке деталей на разных типах станков различными способами обработки.

Двухфакторный дисперсионный анализ с повторениями

ИТОГИ                      Способ 1      Способ 2      Способ 3      Способ 4      Итого

*Станок 1*

Счет	3	3	3	3	12
Сумма	62,7	61,0	54,8	51,7	230,2
Среднее	20,9	20,3	18,3	17,2	19,2
Дисперсия	0,5	0,3	2,8	0,3	3,1

*Станок 2*

Счет	3	3	3	3	12
Сумма	38,3	38,5	38,7	41,2	156,7
Среднее	12,8	12,8	12,9	13,7	13,1
Дисперсия	1,5	1,2	0,7	0,1	0,8

*Станок 3*

Счет	3	3	3	3	12
Сумма	38,8	43,4	37,9	39,1	159,2
Среднее	12,9	14,5	12,6	13,0	13,3
Дисперсия	0,8	1,0	0,2	0,4	1,0

*Станок 4*

Счет	3	3	3	3	12
Сумма	40,3	42,6	41,1	40,6	164,6
Среднее	13,4	14,2	13,7	13,5	13,7
Дисперсия	0,3	1,1	0,2	1,9	0,7

*Итого*

Счет	12	12	12	12	
Сумма	180,1	185,5	172,5	172,6	
Среднее	15,0	15,5	14,4	14,4	
Дисперсия	13,3	9,7	6,4	3,5	

Дисперсионный анализ

<i>Источник вариации</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-Знач</i>	<i>F крит</i>
Выборка	309,26	3,00	103,09	123,64	0,00	2,90
Столбцы	9,97	3,00	3,32	3,99	0,02	2,90
Взаимодействие	25,68	9,00	2,85	3,42	0,00	2,19
Внутри	26,68	32,00	0,83			
Итого	371,59	47,00				

В строке “Выборка” указаны расчетные значения показателей для фактора А – тип станка. Как видно, расчетное значение F-критерия попадает в кри-

тическую область ( $123,64 > 2,9$ ), подтверждая, что тип станка влияет на выработку деталей. Выборочный коэффициент детерминации для фактора А:

$$R^2 = \frac{309,26}{371,59} = 0,83$$

показывает, что 83% общей выборочной вариации выработки деталей обусловлено влиянием типа станка.

В строке «Столбцы» указаны расчетные значения показателей для фактора В – способ обработки заготовок. Расчетное значение F-критерия также подтверждает, что способ обработки также влияет на выработку деталей ( $3,99 > 2,90$ ). Выборочный коэффициент детерминации для фактора В:

$$R^2 = \frac{9,97}{371,59} = 0,03$$

показывает, что только 3% общей выборочной вариации выработки деталей связано с влиянием способа обработки заготовок.

Значимость фактора взаимодействия ( $3,42 > 2,19$ ) свидетельствует, что эффективность различных типов станков изменяется в зависимости от способа обработки заготовок.

### 3. СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ВЗАИМОСВЯЗЕЙ

#### 3.1. Стандартные статистические функции

В мастере функций есть ряд статистических функций, связанных с режимами «Ковариация» и «Корреляция».

**Функция КОВАР** рассчитывает значение ковариации между двумя массивами данных.

Синтаксис: КОВАР (массив1; массив2)

**Функция КОРЕЛЛ** рассчитывает линейный коэффициент корреляции между массивами данных.

Синтаксис: КОРЕЛЛ (массив1; массив2)

В рассмотренном примере (см. рис. 3.1.1) линейный коэффициент корреляции между динамикой розничного товарооборота и реальных располагаемых денежных доходов составил 0,447.

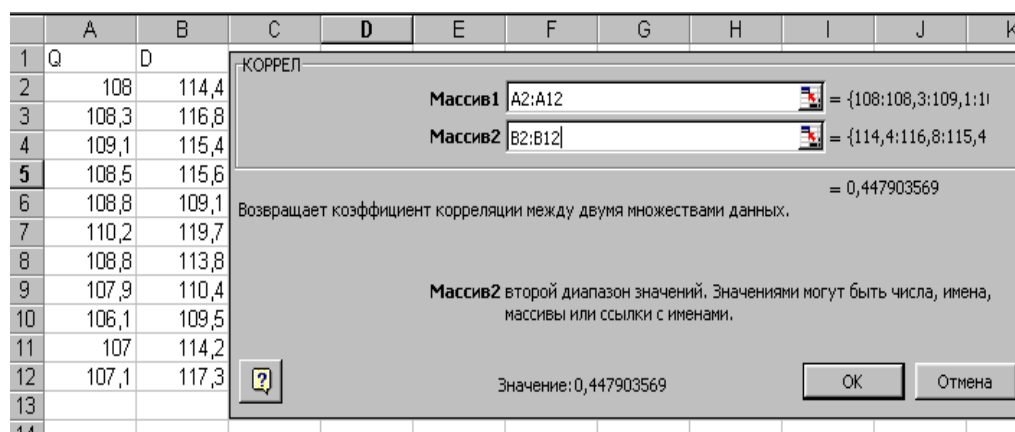


Рисунок 3.1.1. Диалоговое окно функции КОРЕЛЛ

Рассмотрим порядок работы со встроенной статистической **функцией ЛИНЕЙН**, которая определяет коэффициенты парной линейной регрессии:

1) введите исходные данные в блок **A2:B17** или откройте существующий файл с исходными данными;

2) выделите область пустых ячеек **5 x 2** (**5** строк, **2** столбца), например, в блоке **D3 : E7** (см. рис. 20) для вывода результатов регрессионной статистики;

3) в главном меню выберите **Вставка/Функция** (или на панели инструментов щелкните по кнопке **Вставка функции**).

4) В окне **Категория** Мастера функций выберите **Статистические**, в окне **Функция** – **ЛИНЕЙН**. Щелкните по кнопке **OK**.

5) заполните аргументы функции (см. рис. 3.1.2)

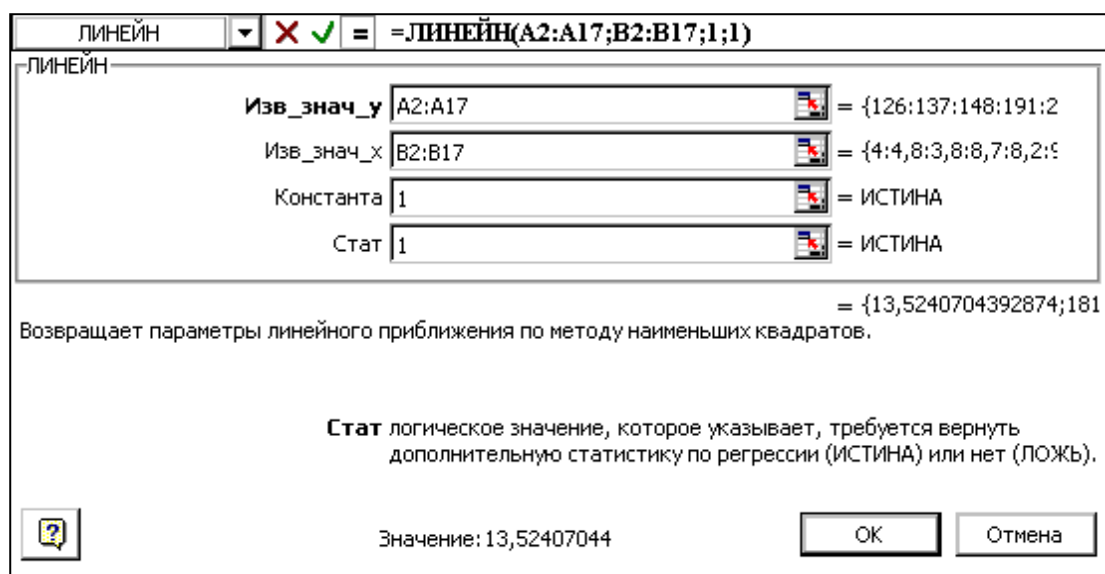


Рисунок 3.1.2. Диалоговое окно ввода аргументов функции ЛИНЕЙН

*Известные значения y* – ссылка на первый столбец блока A2:B17, содержащий данные результативного признака. *Известные значения x* – ссылка на второй столбец этого же блока, содержащий данные независимого признака. *Константа* – логическое значение. Если Константа = 1, то свободный коэффициент *a* рассчитывается обычным образом. Если Константа = 0, то свободный коэффициент *a* равен 0; в данном примере укажите 1. *Статистика (Стат)* – логическое значение. Если Статистика = 1, то дополнительная информация выводится. Если Статистика = 0, то выводятся только оценки коэффициентов уравнения; мы введем 1.

б) чтобы результат регрессии разместился как массив значений, расположенный в выделенной ранее области D3:E7, после ввода всех аргументов функции ЛИНЕЙН надо одновременно нажать комбинацию клавиш <CTRL>+<SHIFT>+<ENTER> (см. рис. 3.1.3).

E6		= {=ЛИНЕЙН(A2:A17;B2:B17;1;1)}					
	A	B	C	D	E	F	G
1	Y	X					
2	126	4					
3	137	4,8		13,52407	181,1232		
4	148	3,8		4,271905	44,60038		
5	191	8,7		0,417211	81,28408		
6	274	8,2		10,0224	14		
7	370	9,7		66219	92499,43		
8	432	14,7					
9	445	18,7					
10	367	19,8					
11	367	10,6					
12	321	8,6					
13	307	6,5					
14	331	12,6					
15	345	6,5					
16	364	5,8					
17	384	5,7					

Рисунок 3.1.3. Результат вычисления функции ЛИНЕЙН

По результату вычисления функции ЛИНЕЙН запишем уравнение регрессии:

$$y = 181,12 + 13,52 * x$$

(44,60) (4,27)

В скобках указаны стандартные ошибки коэффициентов.

Дополнительная регрессионная статистика будет выводиться в порядке, указанном в следующей ниже схеме:

Левый столбец массива	Правый столбец массива
Значение коэффициента <b>b</b>	Значение коэффициента <b>a</b>
Стандартная ошибка <b>b</b>	Стандартная ошибка <b>a</b>
Коэффициент детерминации <b>R<sup>2</sup></b>	Среднеквадратическое отклонение <b>y</b>
F-критерий	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

Коэффициент детерминации составил 0,4172, то есть 41,72% дисперсии объема реализации обусловлено дисперсией расходов на рекламу. Это не свидетельствует об их очевидной линейной зависимости. При увеличении расходов на рекламу на 1000 рублей, объем реализации возрастает на 13520 рублей. Наблюдаемое значение критерия Фишера составило 10,02.

Для нахождения критического значения критерия Фишера применяется функция **ФРАСПРОБР**. По аналогии с предыдущим, в окне *Категория* Мастера функций выберите **Статистические**, в окне *Функция* – **ФРАСПРОБР**. Щелкните по кнопке ОК(см .рис. 3.1.4).

Заполните аргументы функции:

*Вероятность* – это вероятность, связанная с F-распределением;

*Степени свободы1* - это числитель степеней свободы, **v1** = m.

ФРАСПРОБР

Вероятность 0,05 = 0,05

Степени\_свободы1 1 = 1

Степени\_свободы2 14 = 14

= 4,600110515

Возвращает обратное значение для F-распределения вероятностей: если  $p = \text{ФРАСП}(x, \dots)$ , то  $\text{ФРАСПОБР}(p, \dots) = x$ .

Степени\_свободы2 знаменатель степеней свободы - число от 1 до  $10^{10}$ , исключая  $10^{10}$ .

Значение: 4,600110515

ОК Отмена

Рисунок 3.1.4. Диалоговое окно функции **ФРАСПРОБР**.

*Степени свободы 2* - это знаменатель степеней свободы, **v2** = n-m-1.



Критическое (табличное) значение при 5% уровне значимости для  $v_1=1$  и  $v_2=14$  равно 4,60. Наблюдаемое значение превышает критическое, то есть признается статистическая значимость и надежность полученных оценок. Регрессионная сумма квадратов составила 66219, остаточная сумма квадратов составила 92499, 93.

### 3.2. Настройка «Пакет анализа»

Режим «**Ковариация**» служит для расчета генеральной ковариации на основе выборочных данных.

Ниже приведены показатели динамики розничного товарооборота (Q), реальной заработной платы (W), реальных располагаемых доходов населения (D) и численности безработных (T) в январе-октябре 2003 года, в % к соответствующему периоду предыдущего года. Ковариация характеризует рассеивание величин и линейные связи между ними. Для независимых случайных величин ковариация равна нулю. Если величина мало отличается от своего математического ожидания (почти не случайна), то показатель ковариации будет мал.

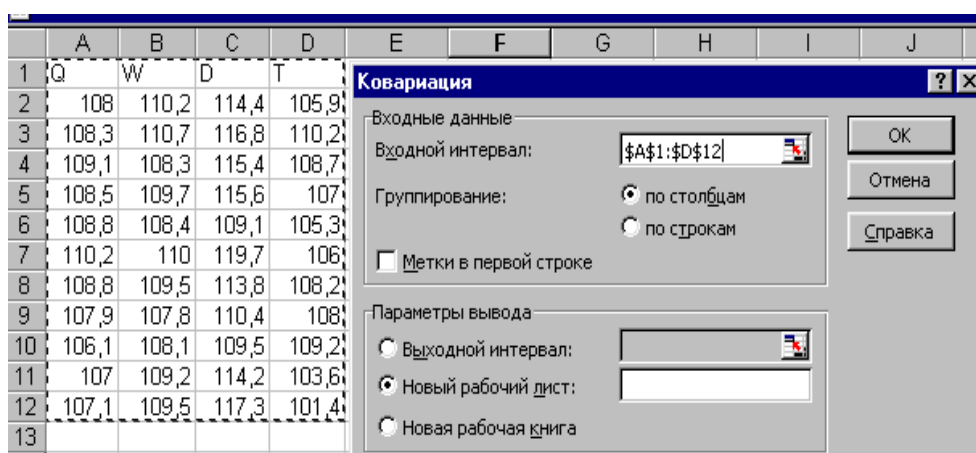


Рисунок 3.2.1. Диалоговое окно режима «Ковариация»

#### Ковариационная матрица

	Q	W	D	T
Q	1,280545			
W	0,299727	0,893636		
D	1,702	2,35	11,276	
T	0,496273	-0,23464	-1,867	6,731636

В данном примере наибольшее рассеивание характерно для динамики реального располагаемого дохода (11,28) и численности безработных (6,73). Прямая линейная связь ярко выражена между динамикой реальной заработной платы и реальными располагаемыми доходами населения (2,35), между динамикой розничного товарооборота и реальными располагаемыми доходами населения (1,70) обратная линейная зависимость наблюдается между реальными располагаемыми доходами населения и численностью безработных (-1,87).

Режим «**Корреляция**» предназначен для расчета генерального и выборочного коэффициентов корреляции соответственно на основе генеральных и выборочных данных (см. рис. 3.2.2.).

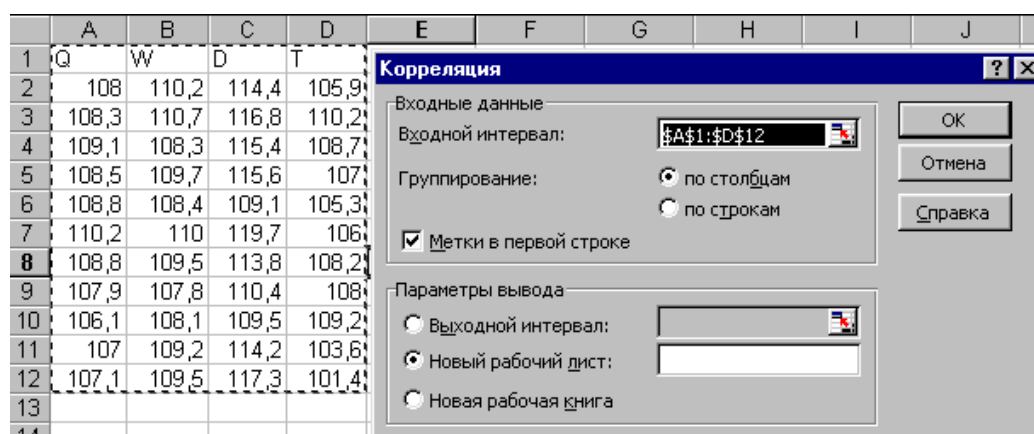


Рисунок 3.2.2. Диалоговое окно режима «Корреляция».

Линейный коэффициент корреляции характеризует тесноту линейной зависимости.

#### Корреляционная матрица

	Q	W	D	T
Q	1			
W	0,280187	1		
D	0,447904	0,740304	1	
T	0,16903	-0,09567	-0,21429	1

Как видно из матрицы, между парами всех показателей существуют стохастические связи. Характер связей состоит в следующем: связь между динамикой реальной заработной платой и реальными располагаемыми доходами является существенной и прямой ( $r_{xy}=0,74$ ); связь между динамикой розничного товарооборота и реальными располагаемыми доходами населения является умеренной и

прямой ( $r_{xy}=0,44$ ); между другими парами показателей имеется слабая линейная связь, причем между динамикой численности безработных, реальной заработной платы и реальными располагаемыми доходами населения связь обратная.

Режим «**Регрессия**» служит для получения оценок коэффициентов линейной регрессии и проверки ее качества. В главном меню выполните **Сервис/Анализ данных**. В панели "Инструменты анализа" выберите **Регрессия**, затем **ОК** и заполните диалоговое окно режима «Регрессия».

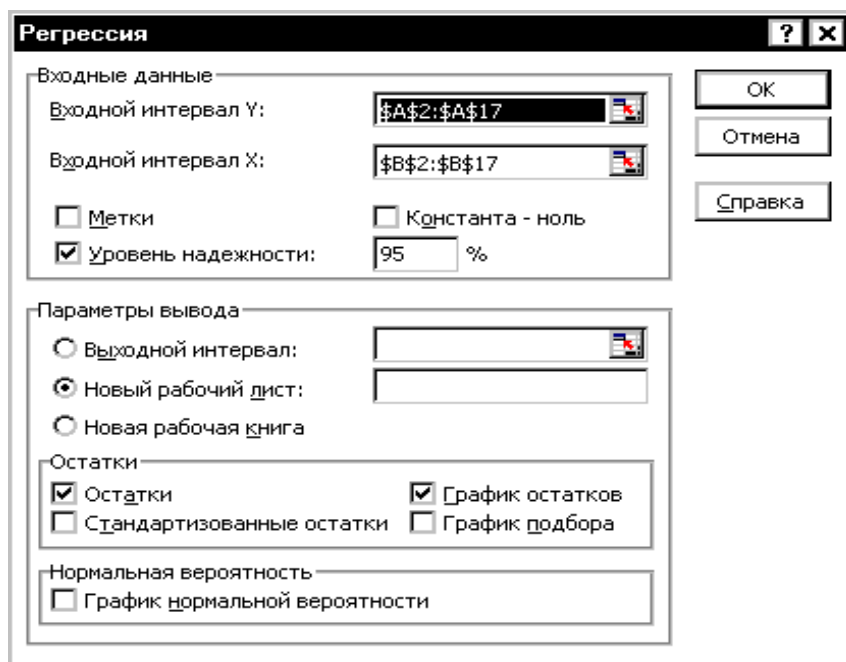


Рисунок 3.2.3. Диалоговое окно режима “Регрессия”

*Входной интервал Y* – диапазон, содержащий данные результативного признака. *Входной интервал X* – диапазон, содержащий данные независимого признака. *Метки* – флажок, указывающий, есть ли в первой строке названия столбцов или нет. *Константа-ноль* – флажок, который указывает на отсутствие свободного коэффициента в уравнении. *Выходной интервал* – указывается левая верхняя ячейка, если вывод результатов производится в заданный интервал. *Новый рабочий лист* - по умолчанию вывод результатов производится на новый рабочий лист, можно указать его имя. Чтобы получить информацию об остатках, необходимо установить соответствующие флажки в этом диалоговом окне. В заключение щелкните по кнопке **ОК**. Рассмотрим пример.

Имеются следующие поквартальные данные по торговой фирме:

Y	126	137	148	191	274	370	432	445	367	367	321	307	331	345	364	384
X	4	4,8	3,8	8,7	8,2	9,7	14,7	18,7	19,8	10,6	8,6	6,5	12,6	6,5	5,8	5,7

Y – объем реализации (тыс. руб), X – расходы на рекламу (тыс. руб).

Оценим коэффициенты уравнения парной линейной регрессии объема реализации и расходов на рекламу. Результаты представлены ниже.

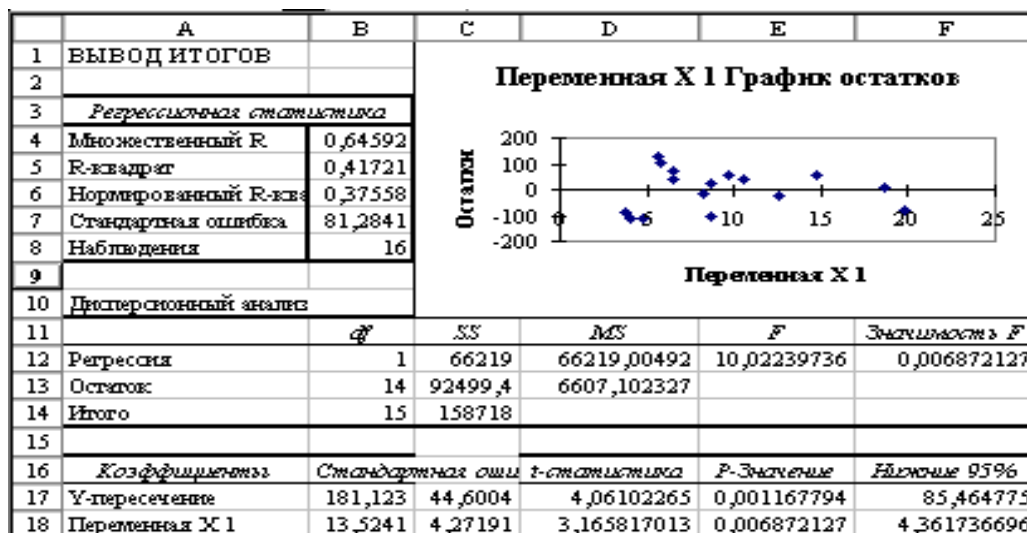


Рисунок 3.2.4. Результаты применения режима Регрессия.

Помимо регрессионной статистики, они содержат таблицу дисперсионного анализа, наблюдаемые значения статистики Стьюдента, доверительные интервалы коэффициентов регрессии, остатки модели, график остатков.

Рассмотрим результаты подробнее.

Регрессионная статистика	
Множественный R	0,64591837- коэффициент парной корреляции
R-квадрат	0,41721054 – коэффициент детерминации
Нормированный R-квадрат	0,37558272 – скорректированный коэф-нт детерминации
Стандартная ошибка	81,2840841 – среднее квадратическое отклонение Y
Наблюдения	16 – количество наблюдений

Коэффициент парной корреляции, равный 0,6459, свидетельствует об умеренной связи между объемом реализации и расходами на рекламу. Скорректированный коэффициент детерминации скорректирован на число степеней свободы, он указывает, что 37,5% дисперсии объема реализации объясняет регрессия с расходами на рекламу. Среднее квадратическое отклонение Y ( $\sigma_y$ ) - средний показатель вариации объема реализации составил 81,28. Таблица

дисперсионного анализа содержит еще один показатель качества регрессии – наблюдаемое значение критерия Фишера. Указана значимость для наблюдаемого значения критерия Фишера – 0,687% (вероятность отвергнуть правильную нулевую гипотезу), она меньше критического уровня значимости, заданного 5 %. В таблице также продемонстрирован расчет факторной и остаточной дисперсий на одну степень свободы.

<b>Дисперсионный анализ</b>					
	<i>Df – число степеней свободы</i>	<i>SS-сумма квадратов отклонений</i>	<i>MS – дисперсия на одну степень свободы</i>	<i>F – наблюдаемое значение критерия Фишера</i>	<i>Значимость F-уровень значимости F</i>
Регрессия	1	66219,00492 (факторная)	66219,00492 (факторная)	10,0223974	0,006872127
Остаток	14	92499,43258 (остаточная)	6607,102327 (остаточная)		
Итого	15	158718,4375 (общая)			

Наблюдаемое значение критерия Стьюдента для коэффициентов  $a$  и  $b$ , а также доверительные интервалы указаны в следующей таблице.

	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>
Y-пересечение	181,1231 (a)	44,6003841	4,06102265	0,001167	85,464775	276,78156
Переменная X 1	13,52407 (b)	4,27190528	3,16581701	0,006872	4,3617366	22,686404

Запишем уравнение регрессии:  $y = 181,12 + 13,52 \cdot x$

(44,60) (4,27) – стандартная ошибка

(4,06) (3,17) – наблюдаемая t-статистика

Для нахождения критического значения критерия Стьюдента применяется статистическая функция СТЬЮДРАСПРОБР. В окне **Категория** Мастера функций выберите **Статистические**, в окне **Функция** – **СТЬЮДРАСПРОБР**, нажмите ОК (см. рис. 3.1.5).

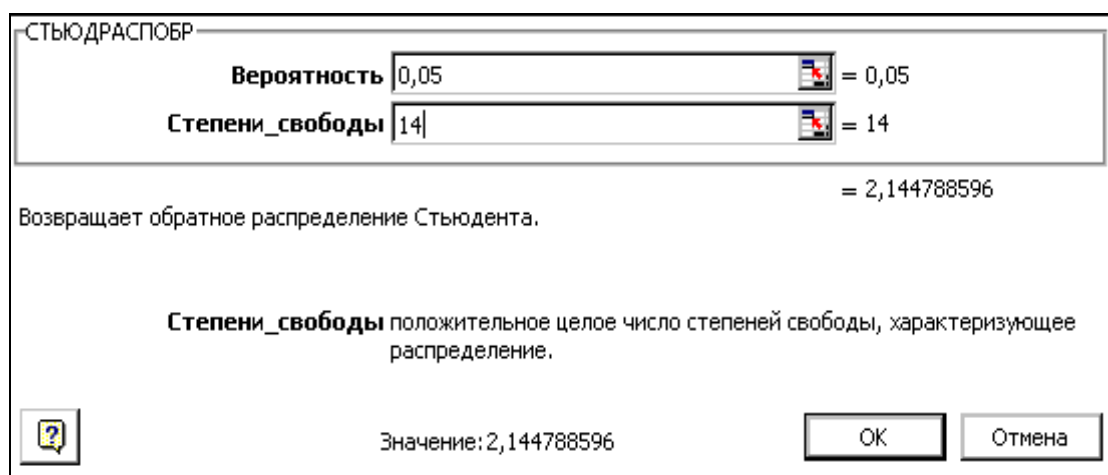


Рисунок 3.1.5. Диалоговое окно функции СТЮДРАСПРОБР после ввода аргументов.

Аргументы функции заполняем, исходя из того, что: поле *вероятность* – это вероятность, соответствующая двустороннему распределению Стьюдента; поле *степени свободы* – это число степеней свободы, характеризующее распределение.

В данном случае критическое (табличное) значение критерия Стьюдента при числе степеней свободы  $\nu = n-2 = 14$  и уровне значимости  $0,05/2=0,025$  составляет **2,1448**. Наблюдаемое значение t-статистики для каждого из коэффициентов превышает критическое. Следовательно, отвергается гипотеза о равенстве коэффициентов нулю и с вероятностью 95% признается их статистическая значимость. Запишем доверительные интервалы, в пределах которых с вероятностью 95% могут находиться значения коэффициентов:

$$85,46 < \mathbf{a} < 276,78; \quad 4,36 < \mathbf{b} < 22,69.$$

## 4. МЕТОДЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

### 4.1.Стандартные статистические функции

В ППП Excel для расчета прогнозируемых значений результативного признака предназначены статистические функции ПРЕДСКАЗ, ТЕНДЕНЦИЯ, РОСТ.

**Функции ПРЕДСКАЗ, ТЕНДЕНЦИЯ** рассчитывают для парной регрессии прогнозируемое значение результативного признака в соответствии с линейным трендом (см. рис.4.1.1) .

Синтаксис: ПРЕДСКАЗ (х; известные значения у; известные значения х)

ТЕНДЕНЦИЯ (известные значения у; известные значения х; новые значения х; конст).

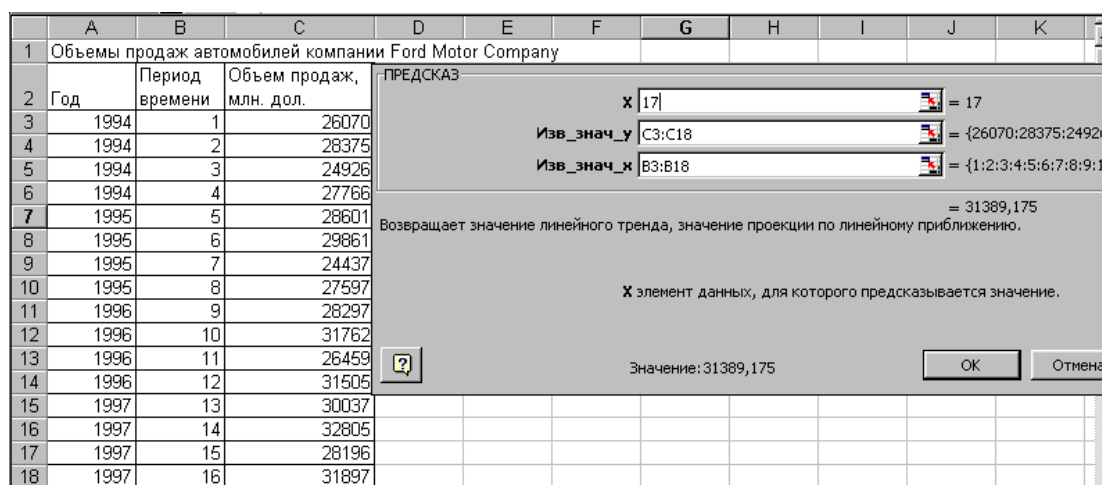


Рисунок 4.1.1. Диалоговое окно статистической функции ПРЕДСКАЗ.

**Функция РОСТ** рассчитывает массив прогнозируемых значений результирующего признака в соответствии с экспоненциальной кривой.



Рисунок 4.1.2. Диалоговое окно статистической функции РОСТ

Синтаксис: РОСТ (известные значения у; известные значения х; новые значения х; конст).

## 4.2. Настройка “Пакет анализа”

Режим “*Скользящее среднее*” служит для сглаживания уровней эмпирического временного ряда на основе метода простой скользящей средней.

	A	B	C	D	E	F	G	H	I	J	K
1	Объемы продаж автомобилей компании Ford Motor Company										
2	Год	Квартал	Объем продаж, млн. дол.	Скользящее среднее							
3	1994	1	26070								
4	1994	2	28375	26457							
5	1994	3	24926	27022,33							
6	1994	4	27766	27097,67							
7	1995	1	28601	28742,67							
8	1995	2	29861	27633							
9	1995	3	24437	27298,33							
10	1995	4	27597	26777							
11	1996	1	28297	29218,67							
12	1996	2	31762	28839,33							
13	1996	3	26459	29908,67							
14	1996	4	31505	29333,67							
15	1997	1	30037	31449							
16	1997	2	32805	30346							
17	1997	3	28196	30966							
18	1997	4	31897								

**Скользящее среднее** ? X

Входные данные  
 Входной интервал:

☒ Метки в первой строке  
 Интервал:

Параметры вывода  
 Выходной интервал:

Новый рабочий лист:

Новая рабочая книга:

☒ Вывод графика ☐ Стандартные погрешности

OK Отмена Справка

Рисунок 4.2.2. Диалоговое окно режима “Скользящее среднее”.

В поле «Интервал» вводится интервал сглаживания (по умолчанию интервал сглаживания равен трем).

В режиме “*Экспоненциальное сглаживание*” реализован метод простого экспоненциального сглаживания.

	A	B	C	D	E	F	G	H	I	J	K
1	Объемы продаж автомобилей компании Ford Motor Company										
2	Год	Квартал	Объем продаж, млн. дол.	Экспон. средняя							
3	1994	1	26070								
4	1994	2	28375	26070							
5	1994	3	24926	27453							
6	1994	4	27766	25936,8							
7	1995	1	28601	27034,32							
8	1995	2	29861	27974,33							
9	1995	3	24437	29106,33							
10	1995	4	27597	26304,73							
11	1996	1	28297	27080,09							
12	1996	2	31762	27810,24							
13	1996	3	26459	30181,29							
14	1996	4	31505	27947,92							
15	1997	1	30037	30082,17							
16	1997	2	32805	30055,07							
17	1997	3	28196	31705,03							
18	1997	4	31897	29599,61							

**Экспоненциальное сглаживание** ? X

Входные данные  
 Входной интервал:

Фактор затухания:

☒ Метки

Параметры вывода  
 Выходной интервал:

Новый рабочий лист:

Новая рабочая книга:

☒ Вывод графика ☐ Стандартные погрешности

OK Отмена Справка

Рисунок 4.2.3. Диалоговое окно режима “Экспоненциальное сглаживание”.

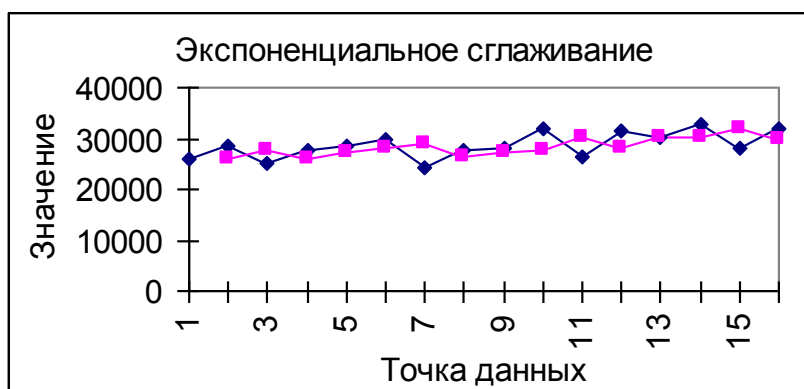


Рисунок 4.2.4. График фактических и теоретических уровней временного ряда.



В поле «Фактор затухания» вводится значение коэффициента экспоненциального сглаживания (от 0 до 1).

Выравнивание временного ряда методом простой скользящей средней и методом экспоненциального сглаживания не позволяют выразить основную тенденцию развития (тренд) через функцию времени. Этого недостатка лишен метод аналитического выравнивания.

Построить линию тренда ППП EXCEL позволяет пункт **Диаграмма** в Главном меню.

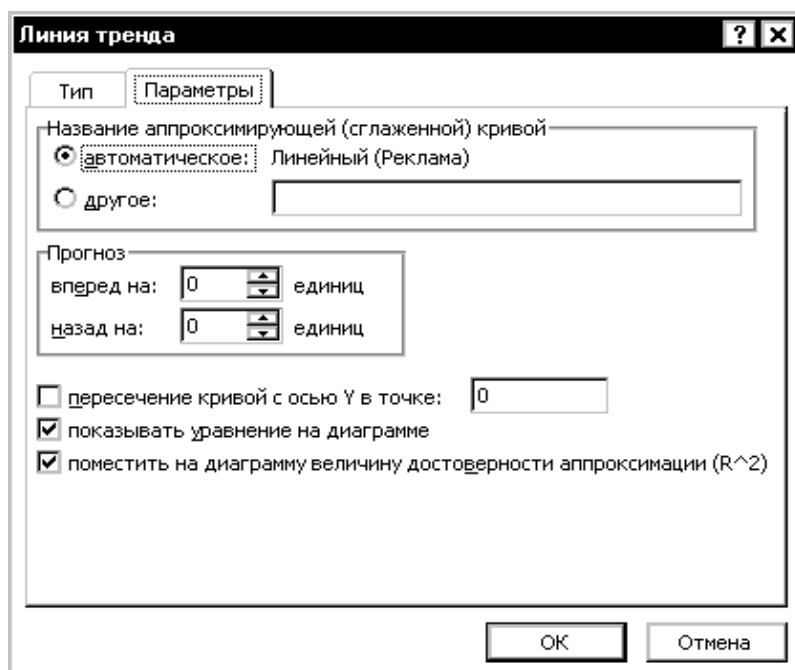


Рисунок 4.2.5. Диалоговое окно Линия тренда, вкладка ПАРАМЕТРЫ .

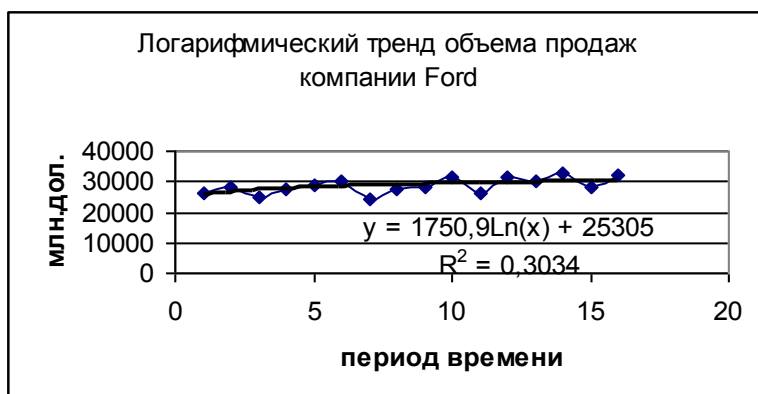
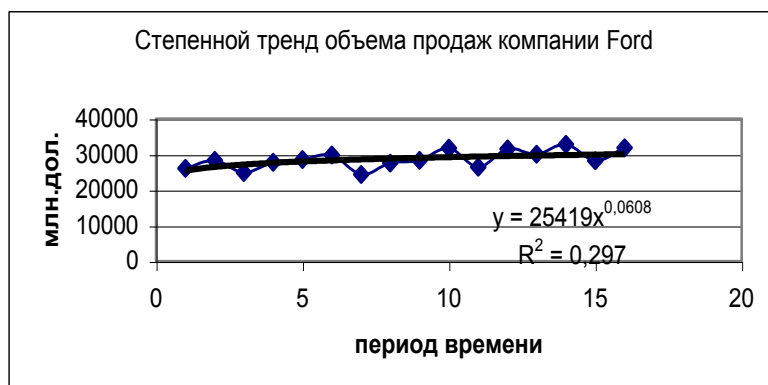
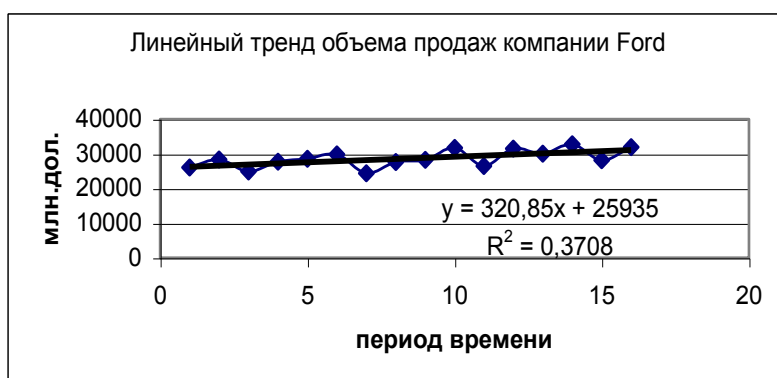
Порядок работы следующий:

1. Введите исходные данные или откройте существующий файл, содержащий анализируемые данные;
2. Для активизации *Мастера диаграмм* в главном меню выберите Вставка/Диаграмма;
3. В окне *Тип* выберите *Точечная*, затем укажите вид точечной диаграммы, щелкните по кнопке *Далее*.
4. Заполните диапазон данных, проверьте, соответствуют ли осям координат данные  $x$  и  $y$ . Если обнаружите несоответствие, то щелкните по кнопке *Ряд* и укажите верный диапазон  $x$  и  $y$ .

5. Заполните параметры диаграммы на разных закладках (названия диаграммы и осей, значения осей и т. п.). Щелкните по кнопке Далее.

6. Укажите место размещения диаграммы. Щелкните -Далее.

Чтобы на точечную диаграмму (поле корреляции) поместить линию регрессии, выделите область построения диаграммы, в главном меню выберите **Диаграмма/Добавить линию тренда**. Выберите тип линии тренда и для отображения на диаграмме уравнения регрессии и значения коэффициента детерминации установите соответствующие флажки на вкладке *Параметры*. (см. рис.4.2.5) Щелкните по кнопке Ок.. Ниже представлены разные типы трендов.



## **Часть 2. Статистический анализ данных в системе STATISTICA**

Интегрированная система статистического анализа и обработки данных STATISTICA состоит из следующих компонент:

- электронных таблиц для ввода исходных данных и специализированных таблиц для вывода численных результатов анализа;
- графической системы для визуализации данных и результатов статистического анализа;
- набора модулей статистических процедур;
- встроенных языков программирования.

Для запуска системы нажмите кнопку Пуск в Windows (левый нижний угол экрана), укажите в меню курсором мыши на команду Программы. В появившемся меню выберите STATISTICA и далее подведите курсор к STATISTICA. На экране появится переключатель модулей с заголовком Statistica Module Switcher. Он содержит перечень всех модулей системы.

Система STATISTICA состоит из набора модулей, в каждом из которых собрана тематически связанная группа процедур. При переключении модулей можно либо оставлять открытым только одно окно системы, либо все вызванные ранее модули.

Быстро переключаться с одного модуля на другой можно, щелкая мышью на их значках на рабочем столе; активизируя соответствующее окно приложения, если оно уже было открыто; выбирая их в меню Статистика или в окне Переключатель модулей (щелкая правой кнопкой мыши по серому полю рабочего окна).

STATISTICA включает в себя следующие специализированные статистические модули: Основные статистики и таблицы (Basic Statistics/ Tables), Непараметрическая статистика (Nonparametrics/Dictrib.), Дисперсионный анализ (ANOVA/MANOVA), Множественная регрессия (Multiple Regression), Нелинейное оценивание (Nonlinear Estimation), Кластерный анализ (Cluster Analysis), Факторный анализ (Factor Analysis), Анализ временных рядов и прогнозирование (Time Series/Forecasting), Организация хранения и обработки дан-

ных (Data Management), Канонический анализ (Canonical Analysis), Multidimensional Scaling (Многомерное шкалирование), Дерево классификации (Classification Trees), Корреспондентский анализ (Correspondence Analysis), Структурное моделирование (SEPATH), Анализ надежности (Reliability/Item Analysis), Дискриминантный анализ (Diskriminant Analysis), Лог- линейный анализ (Log-linear Analysis), Анализ выживания (Survival Analysis), Обобщенная линейная модель (General Linear Model), Обобщенная пошаговая регрессия (General Stepwise Regr.), Универсальная линейная модель (Generalized Linear Model), Частные наименьшие квадраты (Partial Least Squares), Компоненты изменения (Variance Components).

Рабочее окно модулей имеет структуру, стандартную для Windows. Верхний заголовок содержит название модуля. Вторая строка – это строка меню, затем панель инструментов и рабочая область. Меню каждого модуля содержит систему выпадающих меню и построено как меню приложений Windows: File (операции с файлами), Edit (операции по редактированию файлов), View (изменение внешнего вида панели инструментов), Analysis (переключатель режимов модуля – специфичен для Statistica) , Graphs (построение графиков), Options (настройка постоянного вида панели инструментов), Window (окна), Help (помощь).

## 1. ОРГАНИЗАЦИЯ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ В СИСТЕМЕ STATISTICA – Модуль Data Management

Покажем, как создаются файлы данных в STATISTICA. Исходное положение: вы находитесь в переключателе модулей.

В таблице приведены данные о тарифах на рекламу в газете «Известия»:

Длина (мм)	Ширина (мм)	Площадь	Цена (долл.)
1	2	3	4
378	517	195426	21500
187	508	94996	11075
187	254	47498	5705

1	2	3	4
92	254	23368	2940
92	127	11684	1515
44	127	5588	780
44	60	2640	405

В переключателе модулей (см. рис. 1.1) выберем модуль Data Management и нажмем кнопку Switch To (Перейти на). В рабочем окне имеется пустая электронная таблица размером 10 x 10 (10 переменных с именами VAR1, VAR2,... VAR10 и 10 пронумерованных наблюдений-случаев) и переключатель режимов модуля Data Management. Имеется 2 способа получения необходимой нам таблицы размером 4 x 7.

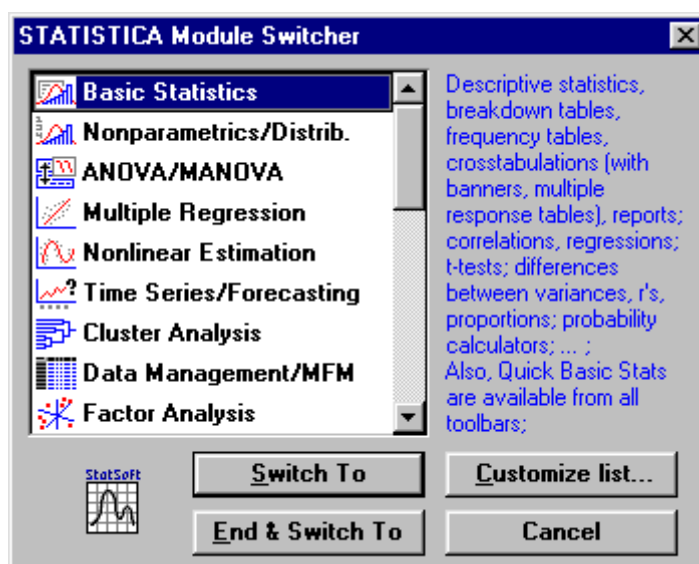


Рисунок 1.1. Переключатель модулей системы STATISTICA.

1 способ. Настройка уже имеющейся таблицы размером 10 x 10. Щелкните правой кнопкой мыши на заголовок 5 столбца (переменной VAR5), в открывшемся диалоговом окне редактирования переменной выберите Modify Variable (изменение переменной)/ Delete и, согласно подсказкам диалогового окна, удалите переменные с 5 по 10. Аналогичные действия выполните с 8-10 строками - Case (наблюдениями). Сохраните файл с именем: reklama1.sta.

2 способ. В меню Analysis щелкните на Startup Panel (панель запуска) и выберите команду Create new data file (Создать новый файл данных):

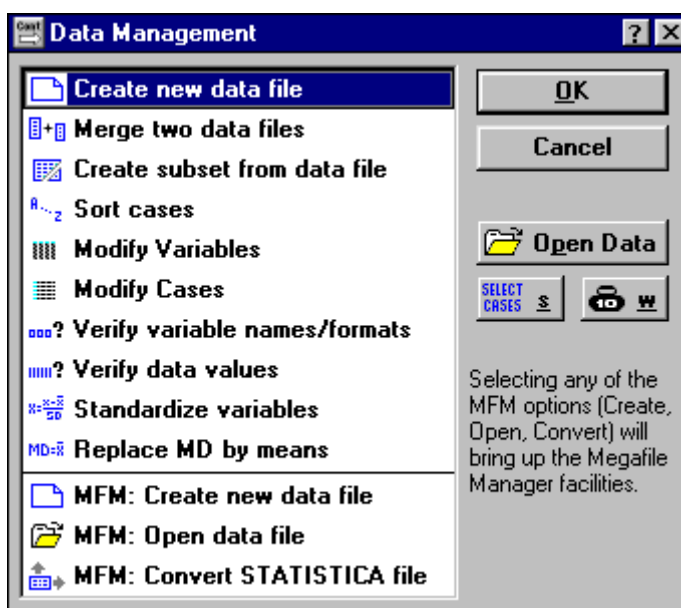


Рисунок 1.2. Переключатель режимов модуля Data Management

В появившемся диалоговом окне укажите количество столбцов (переменных) и строк (наблюдений), запишите имя файла `reklama1.sta` и сохраните файл в папке данных: `C: Statistica/Examples`.

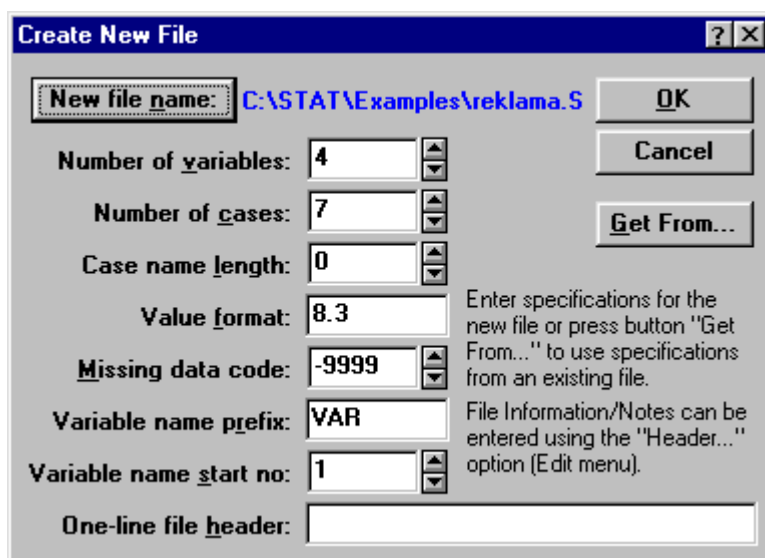


Рисунок 1.3. Диалоговое окно режима Создание нового файла.

При подготовке таблицы к вводу данных требуется указать имена переменных, их тип. Чтобы редактировать отдельную переменную, дважды левой кнопкой мыши щелкните по заголовку переменной и укажите требуемые поля диалогового окна:

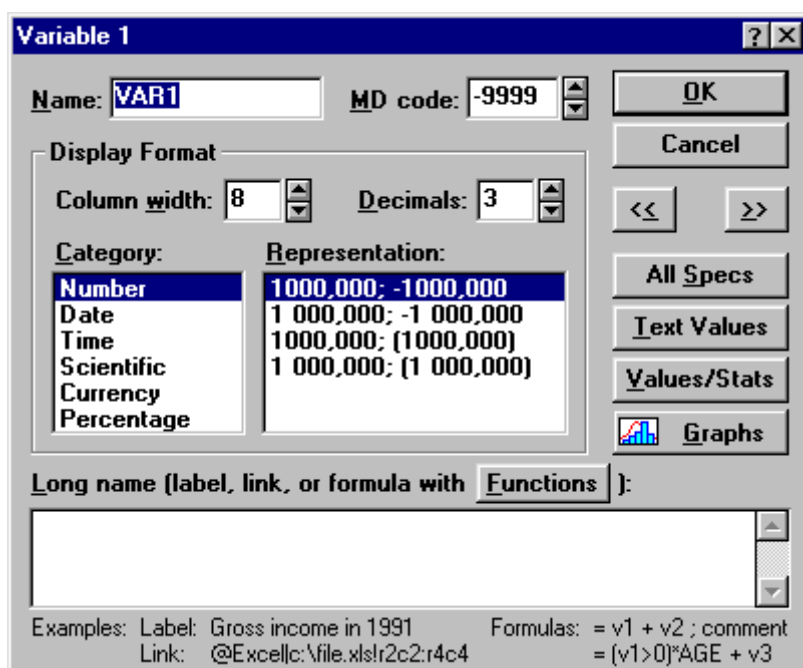


Рисунок 1.4. Диалоговое окно спецификации переменной.

Переместите курсор на белое поле под слова Data: reklama1.sta 4v\* \*7c и дважды щелкните левой кнопкой мыши. В диалоговом окне Data File Header, Notes and Workbook Info (заголовок файла данных, примечания и информация рабочей области) в строке One line Data File Header (одна строка заголовка файла данных) укажите заголовок таблицы. Щелкните Ок. Теперь файл готов для ввода исходных данных. Введите исходные данные или скопируйте их из другого приложения (системы).

Data: reklama1.STA 4v * 7c			
Цена рекламы			
1	2	3	4
ДЛИНА	ШИРИНА	ПЛОЩАДЬ	ЦЕНА
378,000	517,000		21500,00
187,000	508,000		11075,00
187,000	254,000		5705,000
92,000	254,000		2940,000
92,000	127,000		1515,000
44,000	127,000		780,000
44,000	60,000		405,000

Рисунок 1.5. Таблица с введенными с клавиатуры данными.

Заполним данными переменную «Площадь». Щелкните дважды левой кнопкой мыши по заголовку переменной и в окне Variable 3 (переменная 3) запишите порядок вычисления переменной ( $=v1*v2$ ):

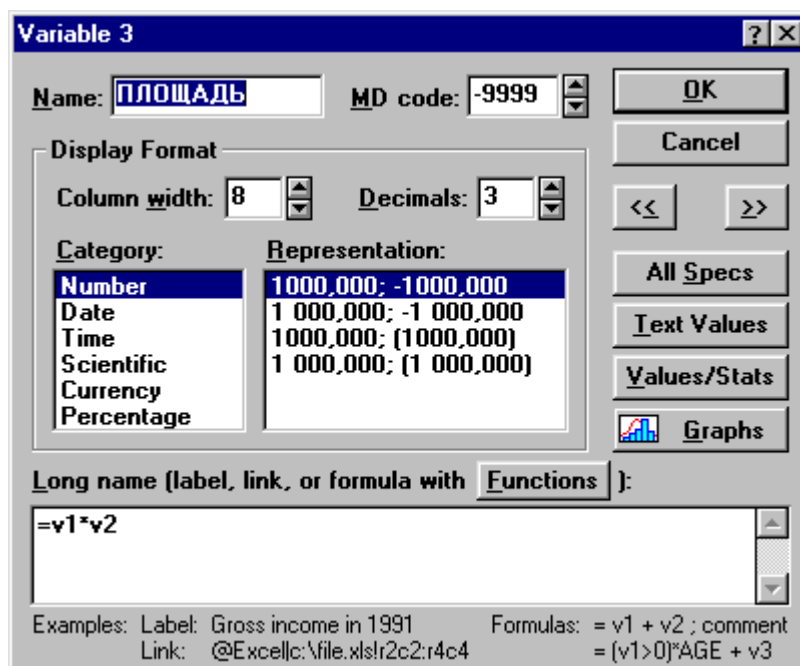


Рисунок 1.6. Вычисление значений переменной «Площадь».

Чтобы сохранить созданный файл в пункте меню File выберите команду Save (сохранить).

## 2. ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ В СИСТЕМЕ STATISTICA – Модуль Basic Statistics/ Tables

В модуле можно определить такие из описательных статистик как среднее значение, выборочную дисперсию, размах вариации, моду, медиану и другие, построить вероятностное распределение (хи-квадрат, Фишера, Стьюдента, Z), таблицы частот. Если вы находитесь в другом модуле, то в пункте меню Analysis выберите команду Quick Basic Stats (быстрые основные статистики). Эта команда имеет выпадающие режимы: Descriptive Statistics (описательная статистика); Correlation matrices (корреляционная матрица); Frequency tables (таблицы частот); Probability Calculator (вероятностный калькулятор); More (другие критерии).

Каждый из режимов реализован в отдельности. Чтобы вывести информацию в комплексе надо включить модуль Basic Statistics в переключателе модулей. Для вызова переключателя модулей в серой части рабочей области активного модуля надо щелкнуть правой кнопкой мыши (см. рис. 2.1).



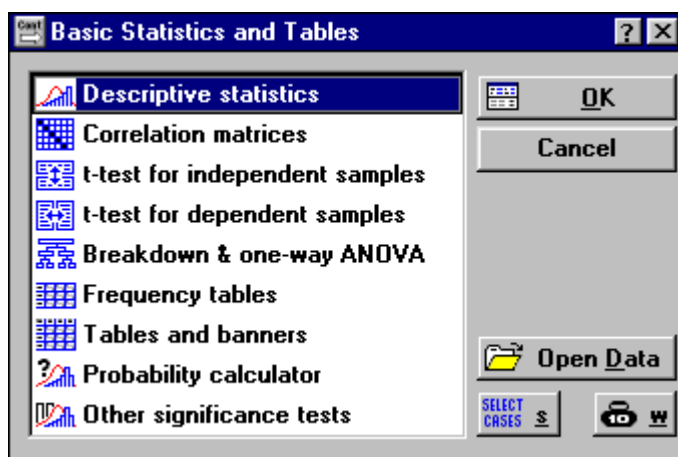


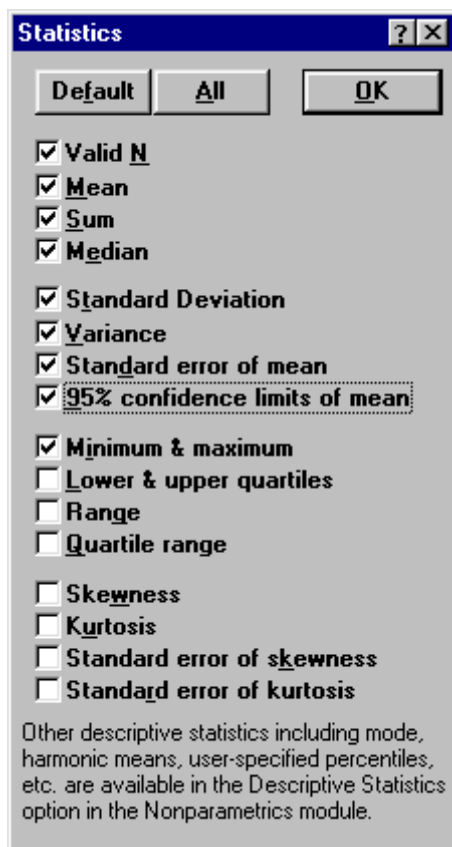
Рисунок 2.1. Стартовая панель модуля Основные статистики и таблицы.

### Вычисление описательных статистик

Щелкнув на кнопку Open Data откройте файл reklama1.sta. На стартовой панели модуля выберите режим Descriptive Statistics (описательная статистика). В диалоговом окне режима на вкладке Var укажите переменную, на вкладке More statistics (другие статистики) - показатели, которые требуется вычислить (см. рис. 2.2).

Рисунок 2.2. Вкладка «Статистики».

Valid N – объем выборки, Mean - выборочное среднее, Sum- сумма, Me-



dian- медиана, Standart Deviation – выборочное стандартное отклонение, Va-

riance- выборочная дисперсия, Standart error of mean – стандартная ошибка среднего, 95% confidence limits of mean – доверительный интервал среднего с вероятностью 95%, Minimum & maximum – минимальное и максимальное значения, Lower & upper quartiles- нижняя и верхняя квартили, Range – размах вариации, Quartile range- квартильный ранг, Skewness- коэффициент асимметрии, Kurtosis- коэффициент эксцесса, Standart error of skewness- стандартная ошибка коэффициента асимметрии, Standart error of kurtosis- стандартная ошибка коэффициента эксцесса.

The image shows two windows of the 'Descriptive Statistics' dialog box in SPSS for the file 'reklama1.sta'. The top window displays basic summary statistics, and the bottom window displays more detailed statistics including confidence intervals and the sum of values.

Continue...	Minimum	Maximum	Range	Variance	Std.Dev.	Standard Error
ДЛИНА	44,0000	378,00	334,00	13971,	118,199	44,675
ШИРИНА	60,0000	517,00	457,00	33816,	183,893	69,505
ЦЕНА	405,0000	21500,00	21095,00	589329E2	7676,775	2901,548

Continue...	Valid N	Mean	Confid. -95,000%	Confid. +95,000%	Median	Sum
ДЛИНА	7	146,286	36,970	255,60	92,000	1024,00
ШИРИНА	7	263,857	93,785	433,93	254,000	1847,00
ЦЕНА	7	6274,286	-825,547	13374,12	2940,000	43920,00

Рисунок 2.3. Таблица с описательными статистиками для переменных.

### Вычисление матрицы парных линейных коэффициентов корреляции

Для продолжения работы в модуле Основные статистики и таблицы в пункте меню Analysis выберите режим Startup panel и перейдите к стартовой панели модуля. На ней выберите Correlation matrices. В диалоговом окне Pearson Product-Moment Correlation на вкладке One variable list (square matrix)- (один список переменных (квадратная матрица)) выделите переменные ЦЕНА и ДЛИНА и щелкните Ок для вывода матрицы с коэффициентами корреляции. Щелкнув мышью по вкладке Correlations диалогового окна Pearson Product-Moment Correlation, можно получить квадратную матрицу коэффициентов корреляции для всех переменных одновременно.

Correlations (reklama1.sta)

Continue... Marked correlations are significant at  $p < .05000$   
N=7 (Casewise deletion of missing data)

Variable	ДЛИНА	ЦЕНА
ДЛИНА	1,00	,97
ЦЕНА	,97	1,00

Рисунок 2.4. Таблица с коэффициентами корреляции.

В этом же диалоговом окне, на вкладке 2D scatterplot (диаграмма рассеяния по двум переменным), можно построить диаграмму рассеяния.

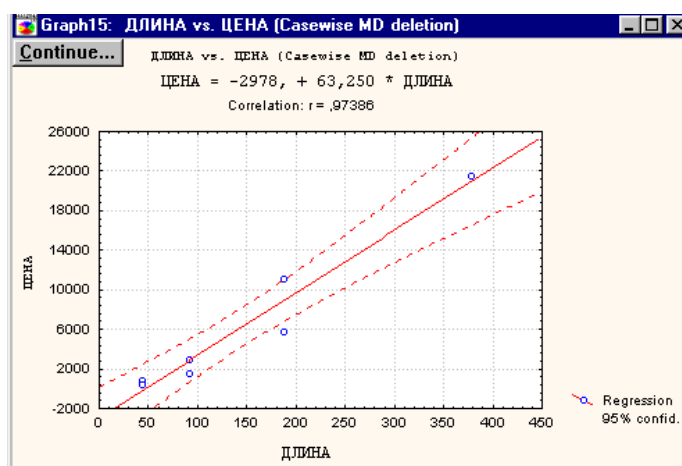


Рисунок 2.5. Диаграмма рассеяния цены на рекламу.

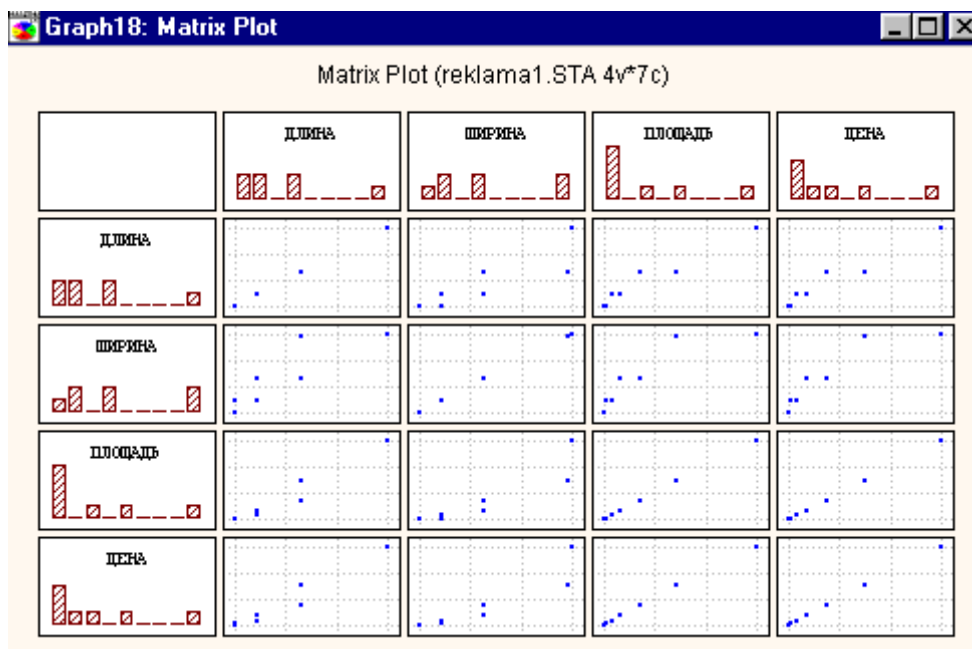


Рисунок 2.6. Матричная диаграмма рассеяния.

Матричная диаграмма рассеяния для группы переменных полезна тем, что позволяет быстро оценить и сравнить распределения выбранных перемен-

ных и форму зависимости (линейная или нелинейная) и направление связи между ними. В пункте меню Graphs (графическая галерея) выберите опцию Quick Statsgraphs (быстрые графики), в ней опцию Matrix scatterplot (матричная диаграмма рассеяния) и Casewise MD deletion.

### 3. ГРАФИЧЕСКИЕ ВОЗМОЖНОСТИ СИСТЕМЫ STATISTICA.

Графики можно построить по таблице исходных данных (статистические графики для первичного анализа исходных данных) и по таблице результатов (пользовательские графики). Графическая галерея Statistica позволяет выбрать сотни различных типов графиков. Диалоговое окно галереи открывается с помощью пункта меню Graphs, который присутствует в каждом модуле.

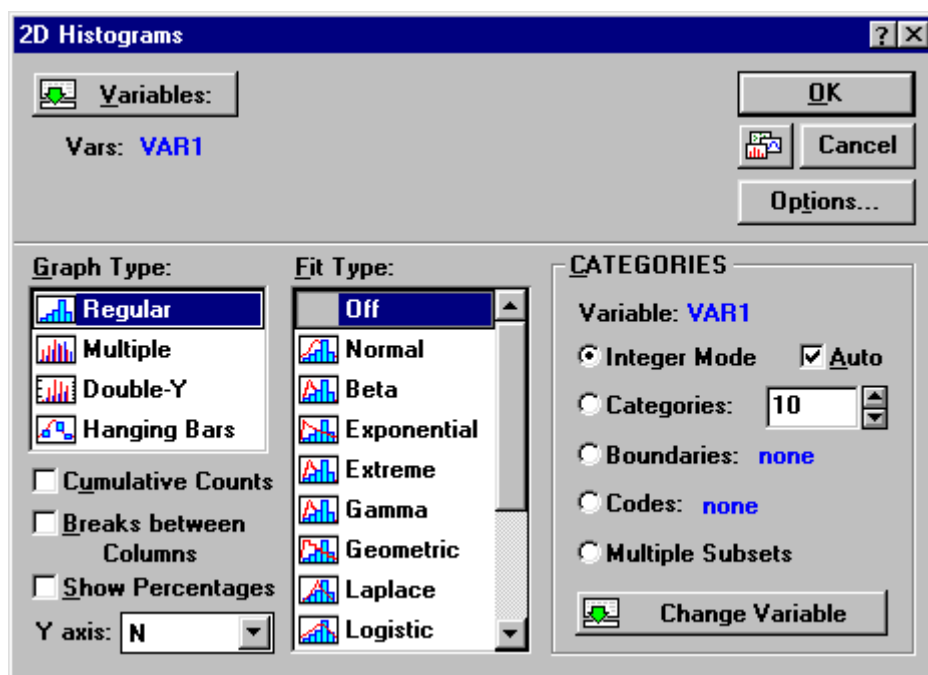


Рисунок 3.1. Диалоговое окно построения гистограммы.

1 шаг. Выбор типа графика. Предположим, вы хотите изучить результаты экзаменов по статистике у студентов вашего потока. Позволяет удобно представить частоту попадания величин (количества студентов) в определенные интервалы (шкала оценок) гистограмма. Создадим файл данных ekz.sta и сохраним его в папке данных C: Statistica/Examples. Оставаясь в модуле Data Management с помощью пункта меню Graphs обратимся к графической галерее и выберем нужную категорию графиков – в нашем случае -, Stats 2D Graphs

(статистические двумерные графики) и в этой группе выберем необходимую группу графиков – Histogram (гистограмма).

2 шаг. Выбор переменных. В диалоговом окне построения гистограммы нажмите на кнопку Variables (переменные) и укажите переменные, которые будут отложены по осям OX и OY.

3 шаг. Построение и сохранение графика. Выберите тип гистораммы Regular (регулярный) и требуемый частый тип off (без сравнения с законом распределения). В разделе Categories (категории) фиксируется переменная, положенная в основу группировки и количество групп (столбцов) в гистограмме, округление интервалов групп до целого (Integer Mode) или расчет интервалов автоматически (Auto), Boundaries (границы интервалов).

.Графики в Statistica хранятся в файлах с расширением \*.stg. Чтобы сохранить график, в пункте меню File выберите команду Save, укажите папку Examples и имя файла.

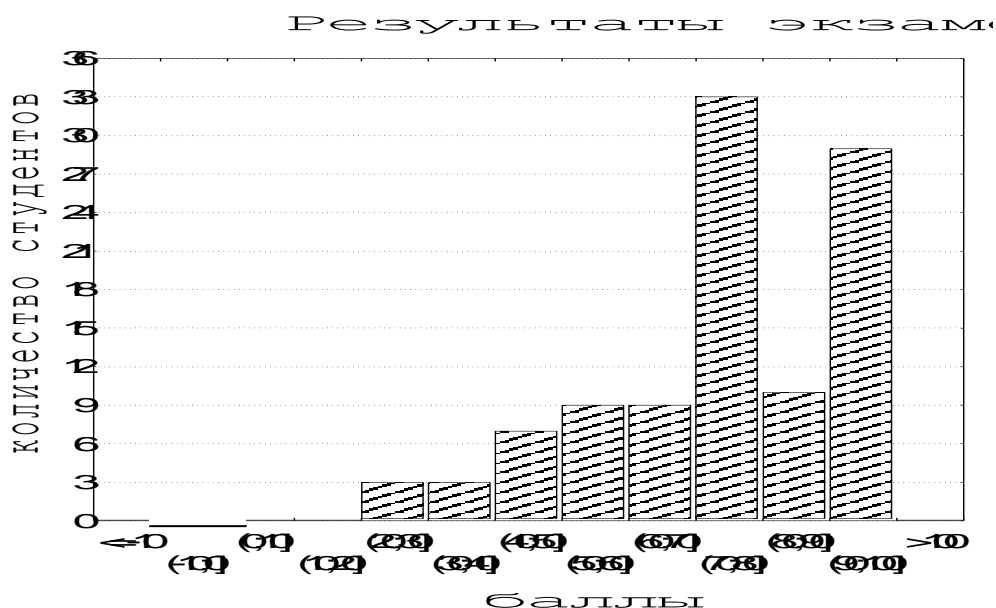


Рисунок 3.2. Гистограмма результатов экзамена по статистике

Упрощенный порядок построения графиков содержится в режиме Quick Stats Graphs пункта меню Graphs. Заранее выделив переменную (или группу переменных), здесь можно быстро построить диаграмму рассеяния, гистограмму, совместив ее с кривой закона распределения, блочные диаграммы. Блочные

диаграммы позволяют анализировать данные на предмет их структуры. Например, построим блочную диаграмму результатов экзамена по статистике.

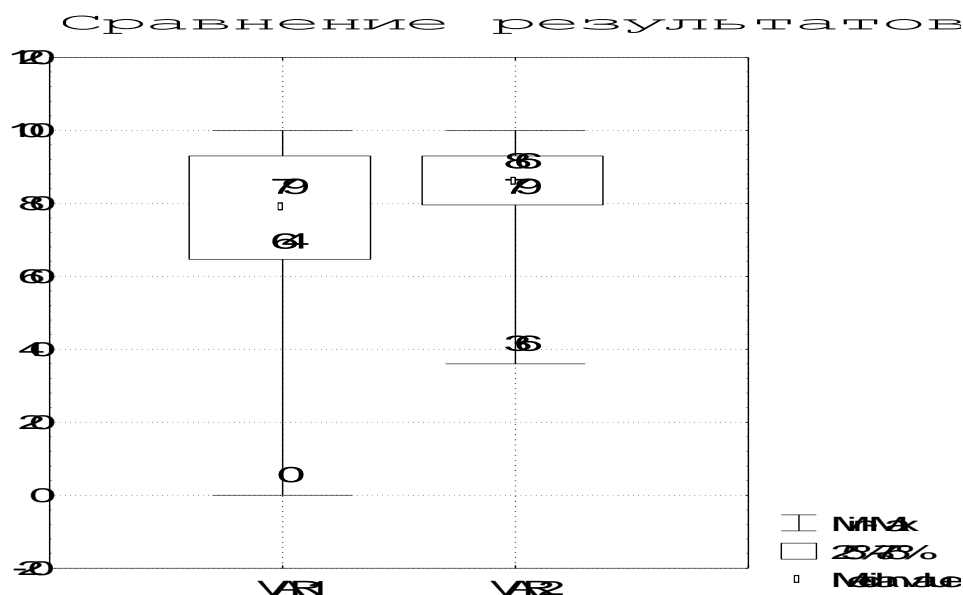


Рисунок 3.3. Блочная диаграмма результатов экзамена по статистике на двух потоках.

Разброс баллов больше на первом потоке, на нем три четверти студентов имеют баллы выше 64, а одна четверть из них выше 93. На втором потоке три четверти студентов имеют баллы выше 86, здесь разброс оценок ниже.

Блочную диаграмму для отдельной переменной можно построить в пункте меню **Grafs**, выбрав в режиме **Quick Stats Graphs** категорию **Box-Whisker of VAR** (график “ящики с усами”).

Чтобы построить блочную диаграмму для группы переменных выполните следующие действия: 1. Откройте модуль **Basic Statistics/Tables** (Основные статистики и таблицы). 2. Выберите в предлагаемом меню строчку **t-test for dependent samples** (t-критерий для зависимых выборок) и нажмите **Ок**. 3. Выберите переменные для анализа. После нажатия кнопки **Variables** (Переменные) в левом списке выберите **VAR1**, в правом – **VAR2**. В строке **Input file** (ввод файла) укажите **Each variable contains the data for one group** (каждая переменная содержит данные для одной группы). 4. Нажмите на кнопку **Box-Whisker plot** и выберите **Median/Quart./Range** (медиана/квартили/размах).

Представим информацию о структуре мужского и женского населения республики (источник: [www.tatstat.ru](http://www.tatstat.ru)) в виде секторных диаграмм. На диаграммах отразим три категории населения: Case 1(13,5%) – моложе трудоспособного возраста, Case 2 (65%)- в трудоспособном возрасте, Case 3 (21,6%)- старше трудоспособного. возраста.

Возрастная структура

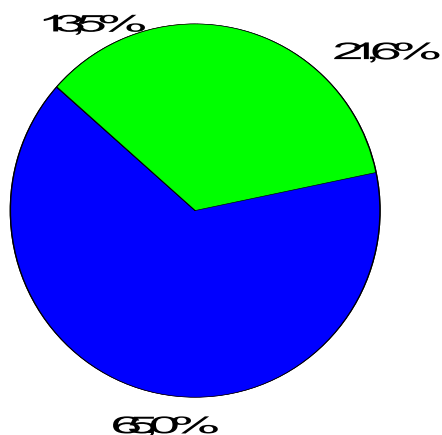


Рисунок 3.4. Секторная диаграмма структуры мужского населения в РТ.

Зависимость объема продаж

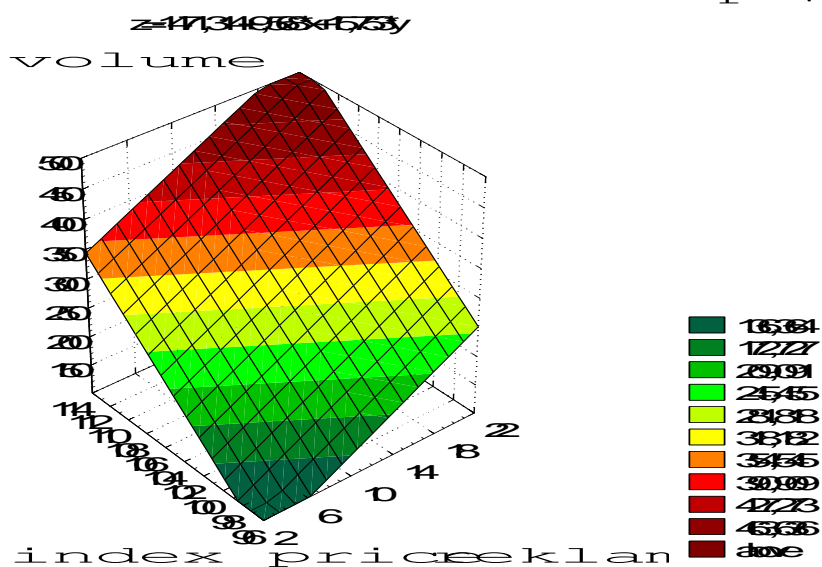


Рисунок 3.5. Пример трехмерного графика.

После того, как график построен, все его структурные компоненты (тип, цвет, вид линии, точек и др.) могут быть настроены пользователем. Доступ к командам настройки реализован при помощи контекстных меню, которые появляются при нажатии на правую кнопку мыши после выделения компонента графика.

#### 4. РЕГРЕССИОННЫЙ АНАЛИЗ В СИСТЕМЕ STATISTICA - модуль Multiple Regression (множественная регрессия)

Данный модуль реализует линейные модели парной и множественной регрессии, содержит блоки дисперсионного анализа, анализа остатков, графическое представление результатов, выполняет расчет показателей общего качества регрессии и статистической значимости оценок.

Создадим файл с данными о курсах валют с 07.04.2004 по 07.05.2004 года и назовем его kurs.sta.

Установим, как курс доллара связан с курсом евро.

В переключателе модулей откройте модуль Multiple Regression (множественная регрессия).

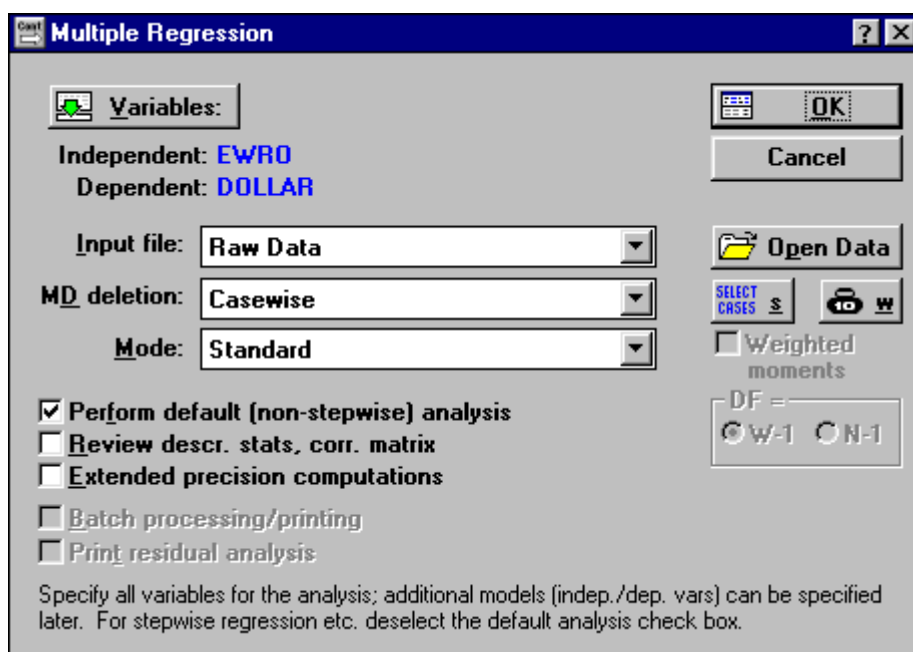


Рисунок 4.1. Стартовая панель модуля Множественная регрессия.



Нажмите кнопку Open Data (открыть данные) и откройте созданный файл данных kurs.sta. Нажмите кнопку Variables (переменные) и в диалоговом окне Select dependent and independent variable list (выбрать списки зависимых и независимых переменных) укажите зависимую и независимую переменную:

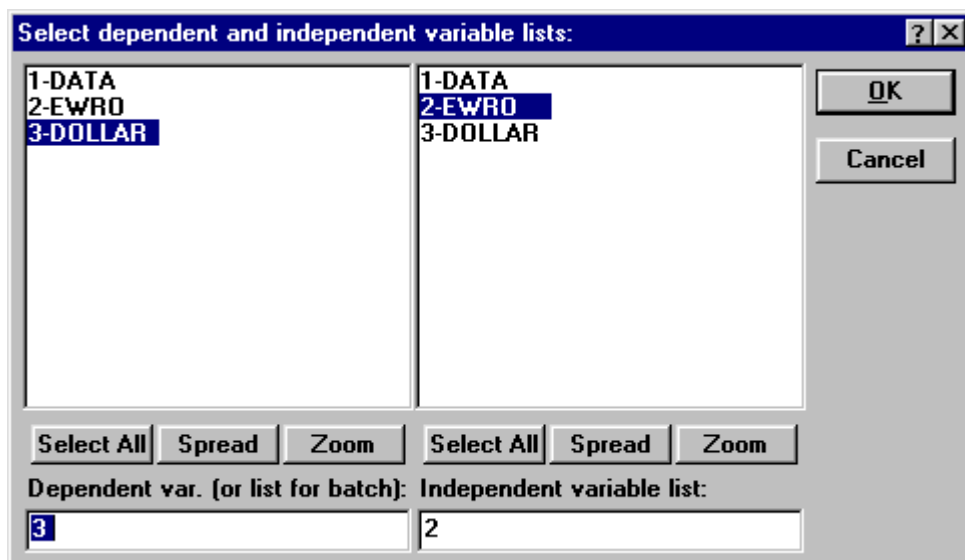


Рисунок 4.2. Окно выбора переменных для анализа.

Выбрав переменные, нажмите Ок и на стартовой панели модуля укажите способ оценивания модели (Mode) стандартный (Standart). Щелкните Ок.

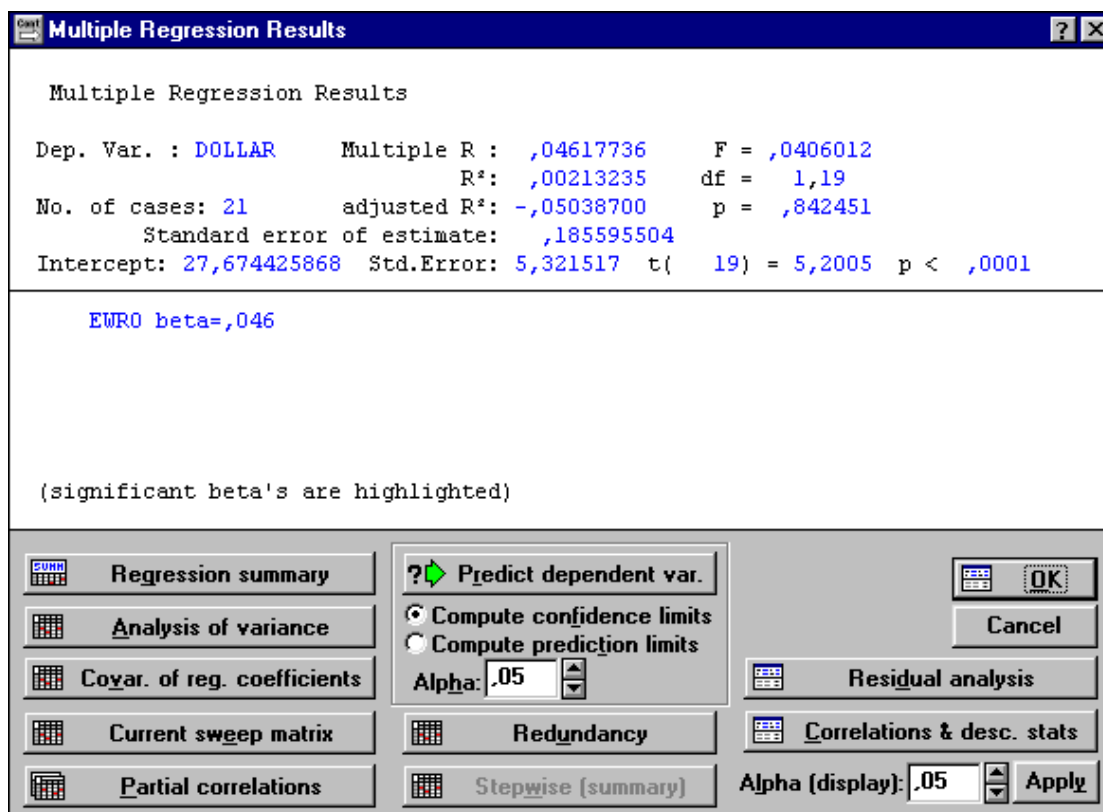


Рисунок 4.3. Окно вывода результатов.

В нем приводятся характеристики и общие показатели качества регрессии: Dep. Var. – зависимая переменная; No. of cases – количество наблюдений; Multiple R: - линейный коэффициент корреляции;  $R^2$  – коэффициент детерминации; Adjusted  $R^2$  – скорректированный коэффициент детерминации; F- критерий Фишера, df- число степеней свободы для критерия, p – уровень значимости для критерия; Standart error of estimate – стандартная ошибка оценки (мера рассеяния наблюдаемых значений относительно регрессионной прямой); Intercept – оценка свободного члена регрессии, Std. Error- стандартная ошибка оценки свободного члена, t(df) and p-value (значение t-критерия и уровень значимости); Beta EWRO- вета-коэффициент перед независимой переменной; Significant beta's are highlighted- значимый бета-коэффициент выделяется красным цветом.

В диалоговом окне имеются кнопки, открывающие другие таблицы результатов: Regression summary- Итоговые оценки регрессии; Analysis of variance – дисперсионный анализ; Covar. of reg. Coefficients –ковариация коэффициентов регрессии; Current sweep matrix – развернутая матрица парных коэффициентов корреляции; Partial correlations- частные коэффициенты корреляции; Redundancy- избыточность; Residual analysis- анализ остатков; Correlations & desc. stats – коэффициент корреляции и описательная статистика; Stepwise (summary) – Итоговый результат пошаговой регрессии; Predict dependent var.- предсказанные значения зависимой переменной и расчет доверительных интервалов.

В нашем примере общие характеристики регрессии свидетельствуют об отсутствии статистической связи между курсами валют в изучаемом периоде времени и о нецелесообразности регрессионного анализа.

Изучим взаимосвязь доходов на одну акцию (зависимая переменная) и курса акций (независимая переменная):

Y	0.24	0.50	0.60	-0.22	-0.81	-0.21	0.21	0.24	-1.00	-0.32
X	17.88	24.75	37.00	11.38	18.75	9.38	17.00	15.00	15.00	5.38

Y	0.02	0.12	-0.87	-0.66	-0.16	-0.57	-0.36	-0.9	-1.1	-0.27
X	11.75	11.38	5.25	6.38	4.63	7.25	4.5	8.75	3.63	1.75

Создадим файл akzia.sta и проведем регрессионный анализ.

Multiple Regression Results					
Multiple Regression Results					
Dep. Var. : PROFIT	Multiple R :	,60212534	F =	10,23773	
	R <sup>2</sup> :	,36255492	df =	1,18	
No. of cases: 20	adjusted R <sup>2</sup> :	,32714131	p =	,004966	
	Standard error of estimate:	,415589985			
Intercept: -,703271863	Std.Error:	,1626898	t( 18) =	-4,323	p < ,0004
COURSE beta=,602					

Рисунок 4.5. Результат парной регрессии со значимыми оценками.

Из основной информации о результатах оценивания очевидно, что между доходами и курсом акций имеется умеренная линейная связь (коэффициент линейной корреляции составляет 60, 2%), в данной выборке наблюдений 36% вариации дохода объясняет разброс курса акций. Оценка свободного члена в уравнении регрессии составляет -0,703 со стандартной ошибкой 0,16, наблюдаемое значение статистики Стьюдента -4, 323 свидетельствует о статистической значимости свободного члена. Наблюдаемое значение критерия Фишера 10,238 выше критического, подтверждает значимость уравнения парной регрессии.

В функциональной части окна результатов нажмем кнопку Regression summary и получим таблицу итоговых результатов оценивания регрессионной модели:

Regression Summary for Dependent Variable: PROFIT						
Continue... R= ,60212534 RI= ,36255492 Adjusted RI= ,32714131 F(1,18)=10,238 p<,00497 Std.Error of estimate: ,41559						
N=20	BETA	St. Err. of BETA	B	St. Err. of B	t(18)	p-level
Intercept			-,703272	,162690	-4,32278	,000410
COURSE	,602125	,188185	,036089	,011279	3,19964	,004966

Рисунок 4.6. Итоговая таблица регрессии.

В первом столбце таблицы оценка стандартизованного бета-коэффициента регрессии, во втором столбце- его стандартная ошибка, в третьем столбце- точечные оценки свободного члена и коэффициента регрессии, далее –их стандартные ошибки, наблюдаемые значения статистики Стьюдента и уровни значимости оценок.

Оцененная модель имеет вид:

$$\text{PROFIT} = -0,7033 + 0,0361 * \text{COURSE}$$

На графике (Graphs/Scatterplot/Linear) исходные данные и теоретическая прямая имеют вид:

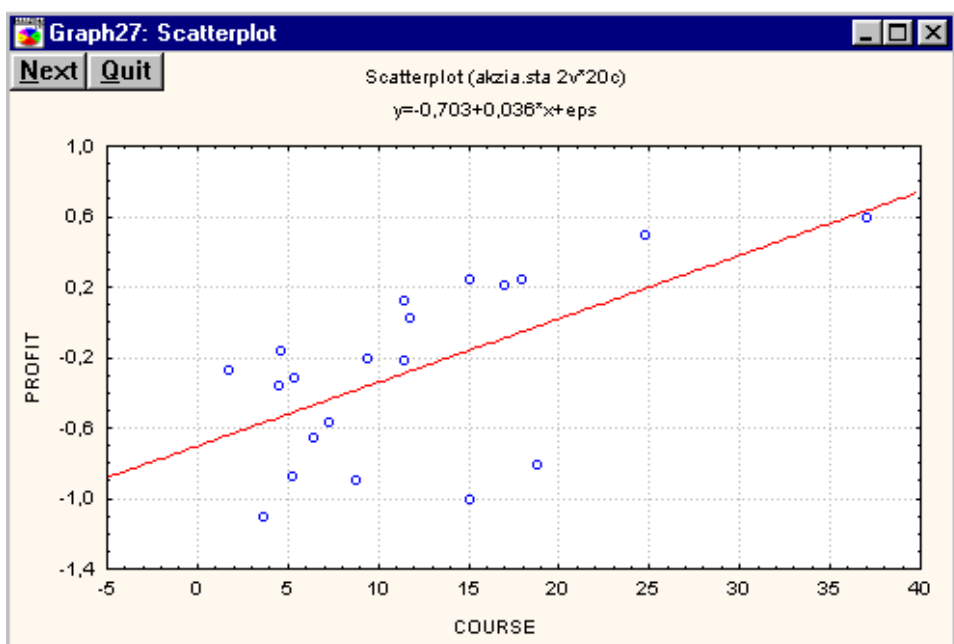


Рисунок 4.7. Линейная регрессия для выборки наблюдений PROFIT и COURSE.

Кнопка Analysis of variance выводит таблицу дисперсионного анализа.

Analysis of Variance; DV: PROFIT (akzia.sta)					
Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress.	1,768209	1	1,768209	10,23773	,004966
Residual	3,108871	18	,172715		
Total	4,877080				

Рисунок 4.8. Таблица дисперсионного анализа.

В первом столбце таблицы записаны суммы квадратов отклонений: регрессионная – 1,76; остаточная – 3,11; общая – 4,88. Во втором столбце – их степени свободы, в третьем – дисперсии (суммы квадратов отклонений в расчете на одну степень свободы), в четвертом столбце- критерий Фишера и уровень значимости его оценки.

Легко можно определить предсказанную величину дохода при заданном курсе акций. Нажмите на кнопку Predict dependent var и в появившемся окне Specify values for independent variables (определить значения независимых переменных) задайте значение независимой переменной, например COURSE=17 и нажмите Ок.

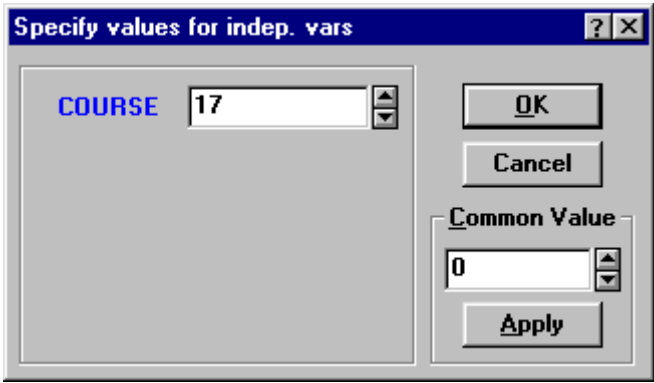
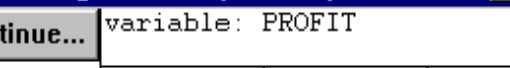


Рисунок 4.9. Окно указания значения независимой переменной.



Predicting Values for (akzia.sta)

**Continue...**

variable: PROFIT

variable	B-Weight	Value	B-Weight * Value
COURSE	,036089	17,00000	,613507
Intercept			-,703272
Predictd			-,089764
-95,0%CL			-,320135
+95,0%CL			,140606

Рисунок 4.10. Предсказанная величина дохода.

В таблице содержится порядок ее расчета и интервальные оценки. Очевидно, что при курсе 17 денежных единиц доход не будет получен (PROFIT=-0,0898).

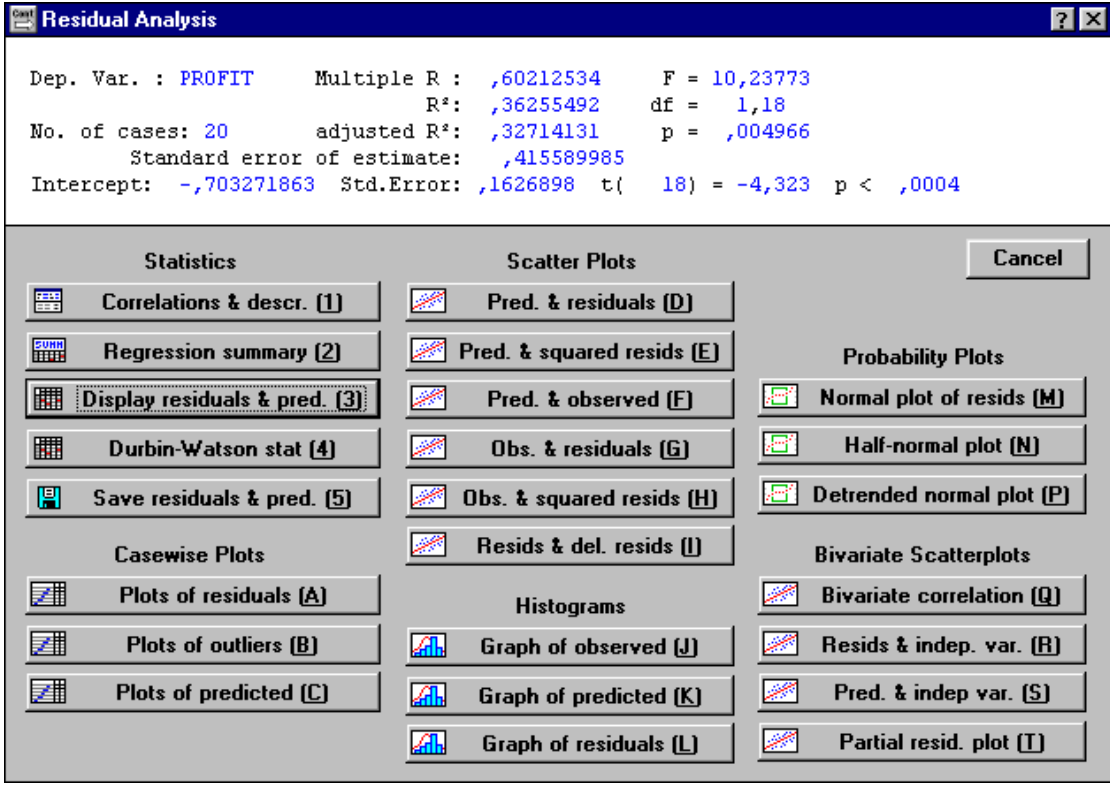


Рисунок 4.11. Диалоговое окно «Анализ остатков».

Анализ адекватности модели основан на анализе остатков. Нажмите кнопку Residual Analysis. В диалоговом окне анализа остатков нажмите на кнопку Obs&residuals (наблюдаемые величины и остатки).

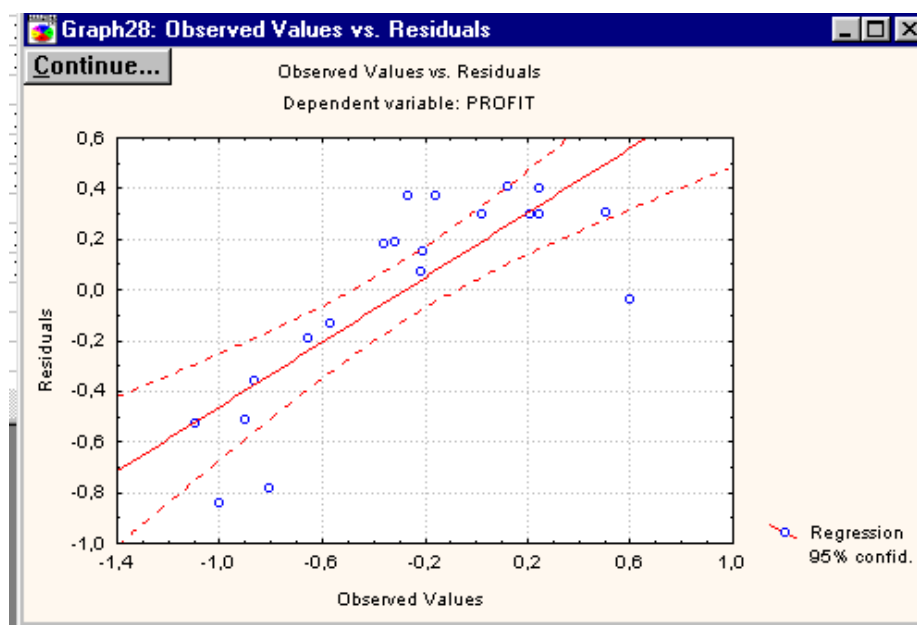


Рисунок 4.12. График «Наблюдаемые переменные – Остатки».

Данный график не свидетельствует о достаточной адекватности модели, поскольку визуально нельзя утверждать о нормальном распределении остатков.

Достаточно часто данные имеют выбросы, которые существенно могут повлиять на построение зависимости. В STATISTICA есть средство, которое позволяет удалять «ненужные» точки или группы точек. Построив график, рис. 23, щелкните по кнопке Кисть (Brush). Справа появится панель Brush. В группе опций Actions (действия) выберите опцию Turn off (выключить), в группе опций Brush выберите опцию Point (Точка)- кисть примет форму точки. Для удаления группы точек кисть может принять форму прямоугольника (опция Rectangle) или произвольной области (опция Lasso). Далее войдите в график и отметьте «ненужную» точку или группу точек. Щелкните на кнопке Update (коррекция) на панели Brushing.

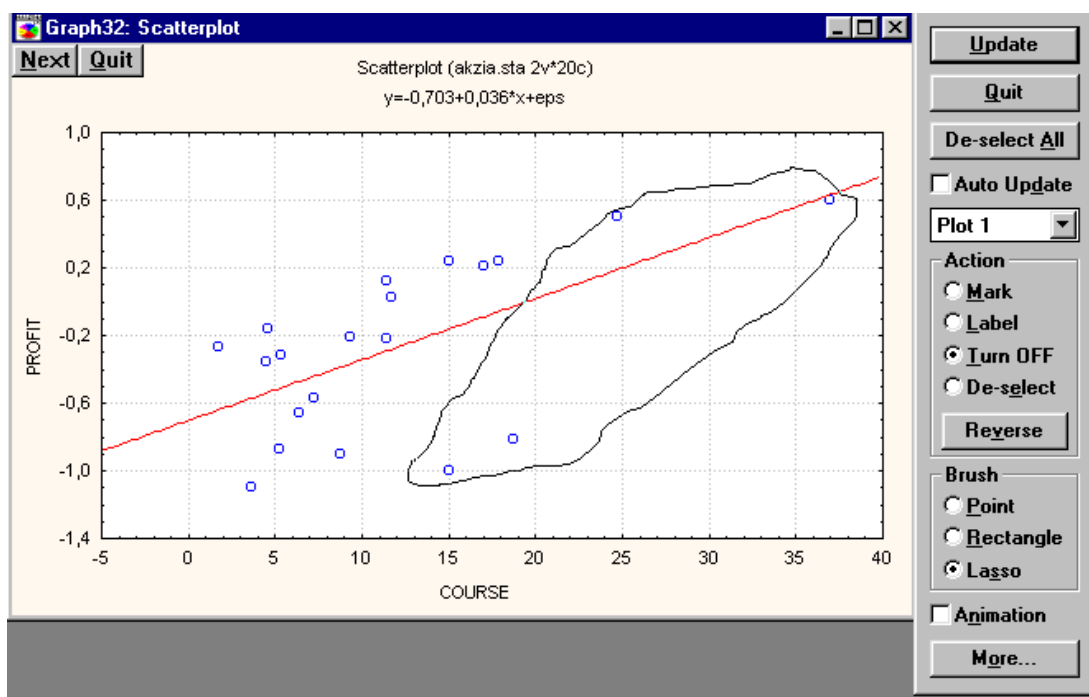


Рисунок 4.13. Панель инструмента Кисть и аномальные точки, заключенные в лассо.

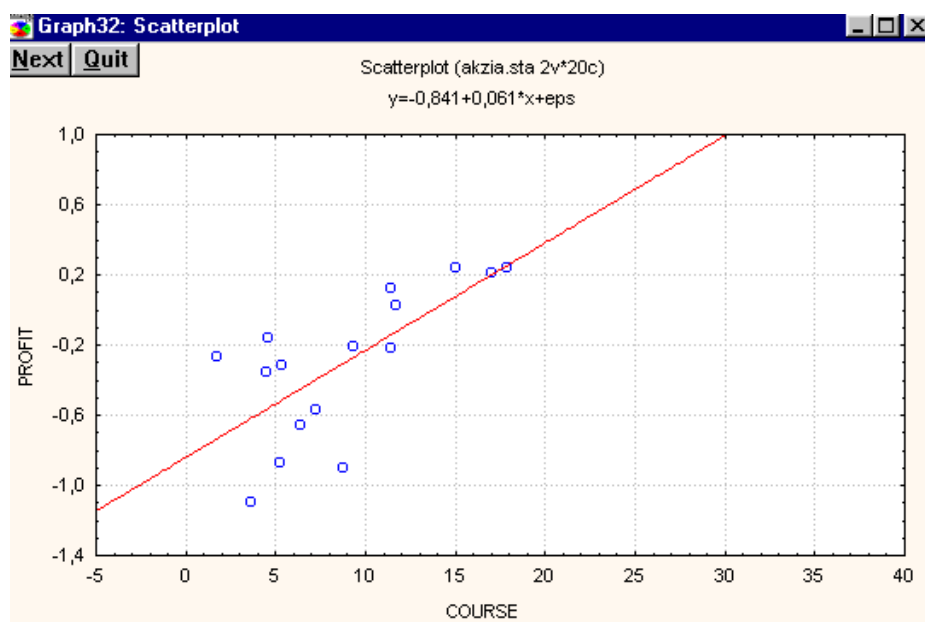


Рисунок 4.14. Данные после удаления аномальных наблюдений и новая регрессионная прямая.

## 5. НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА - модуль Nonparametrics/ Distrib.

Непараметрические методы применяются для анализа малых выборок и для данных, измеренных в малых шкалах. Для оценки степени зависимости ме-

жду переменными рассчитывают ранговые (непараметрические) коэффициенты корреляции. Среди непараметрических процедур в Statistica есть оценка критериев различия для независимых выборок и для зависимых выборок.

Стартовая панель модуля Непараметрические статистики имеет следующий вид:

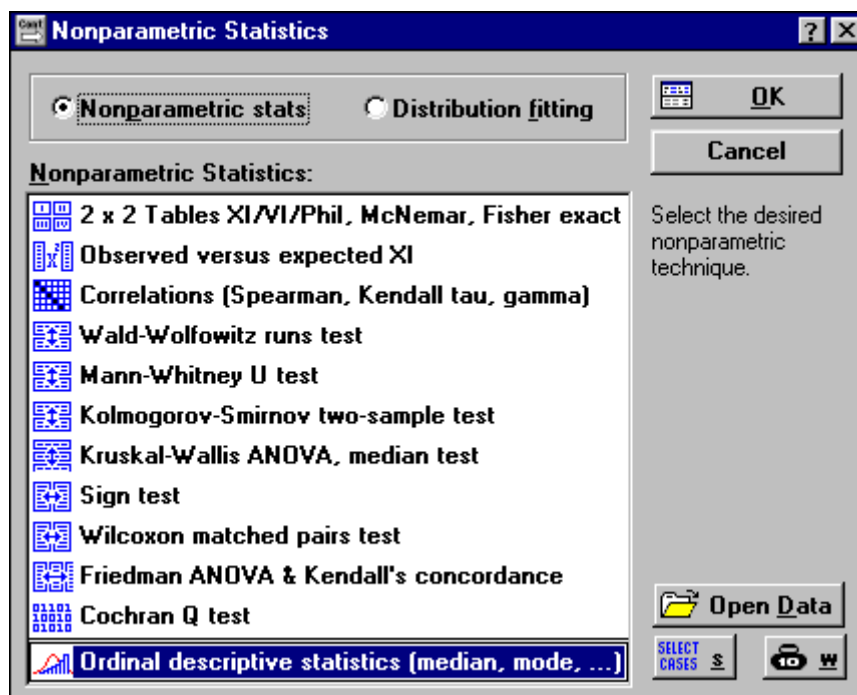


Рисунок 5.1. Стартовая панель модуля Непараметрические статистики.

Опция Correlations (Spearman, Kendall tau, gamma) позволяет вычислить три альтернативы параметрическому коэффициенту Пирсона: коэффициент корреляции Спирмена, коэффициент «тау» Кендалла, коэффициент «гамма». Выясним, зависима ли прибыль двух филиалов в торговой компании, зафиксированная ежемесячно за год.

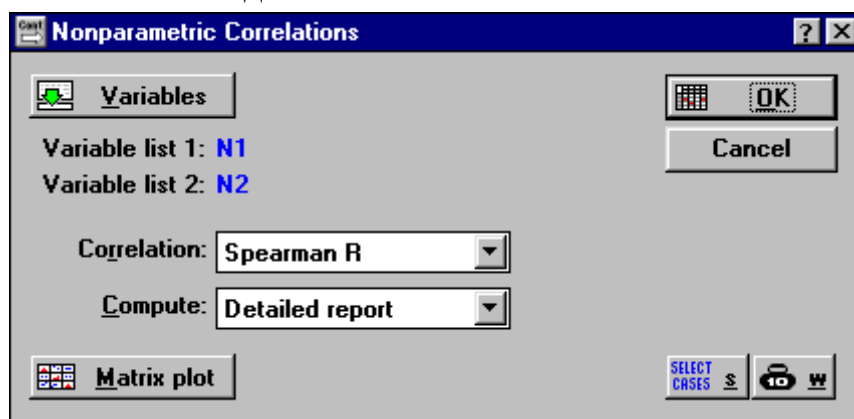


Рисунок 5.2. Диалоговое окно ранговых коэффициентов корреляции.



В диалоговом окне выберем Spearman R и Detailed report (подробный отчет). После нажатия Ок появится результат:

Spearman Rank Order Correlations (firma.sta)				
Continue...		MD pairwise deleted		
Pair of Variables	Valid N	Spearman R	t(N-2)	p-level
N1 & N2	12	,853147	5,171628	,000418

Рисунок 5.3. Расчет коэффициента Спирмена.

Видно, что корреляция между двумя переменными высокосignификантна. Визуализация найденной зависимости возможна двумя способами. Либо нажав кнопку Matrix plot (матричная диаграмма рассеяния), либо щелкнув правой кнопкой мыши по таблице результатов и выбрав опцию Quick Stats Graphs / Scatterplot -диаграмма рассеяния.

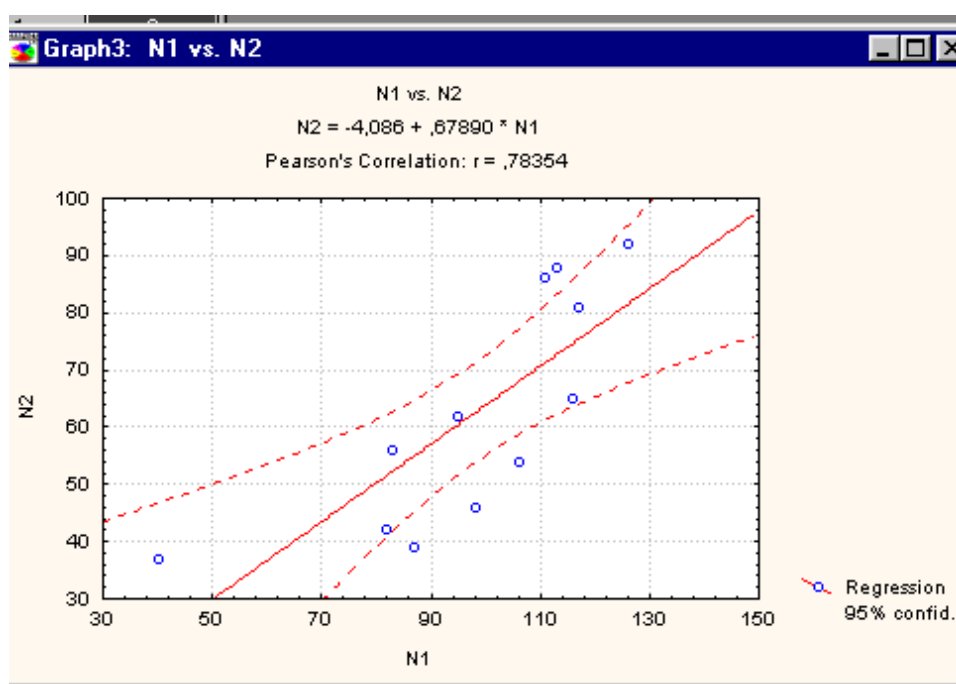


Рисунок 5.4. Диаграмма рассеяния, уравнение зависимости и параметрический коэффициент Пирсона.

Интересно, что корреляция Пирсона меньше корреляции Спирмена. Видимо, рассмотрение рангов (а не самих наблюдений) в действительности улучшает оценку зависимости между переменными, так как подавляет случайную изменчивость и уменьшает воздействие выбросов.

Pair of Variables	Valid N	Kendall Tau	Z	p-level
N1 & N2	12	.696970	3.154337	.001609

Рисунок 5.5. Расчет коэффициента Кендалла.

Статистика Кендалла оценивает разность между вероятностью того, что наблюдаемые значения переменных имеют один и тот же порядок, и вероятностью того, что порядок различный.

На стартовой панели модуля непараметрической статистики также предусмотрена опция Ordinal descriptive statistics (порядковая описательная статистика) для расчета моды, медианы, средней геометрической, средней гармонической, размаха, дисперсии, стандартных ошибок и других оценок описательной статистики. Щелкнув на таблице результатов правой кнопкой мыши, можно построить диаграмму размаха («ящики с усами»).

## 6. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ И ПРОГНОЗИРОВАНИЕ В СИСТЕМЕ STATISTICA – Модуль Time Series/Forecasting

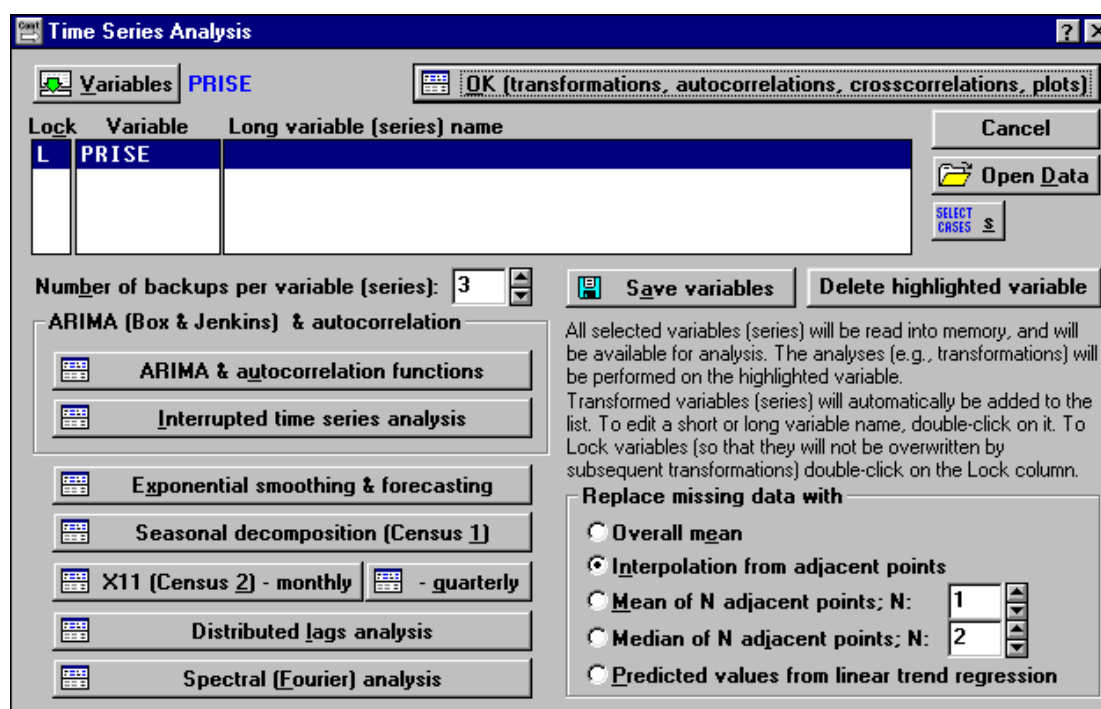


Рисунок 6.1. Стартовая панель модуля Анализ временных рядов

На стартовой панели находятся кнопки методов анализа, которые реализованы в данном модуле: ARIMA – модель авторегрессии и проинтегрированного скользящего среднего (АРПСС); Interrupted time series analysis – анализ прерванного временного ряда (модели интервенции для АРПСС); Exponential smoothing & forecasting – экспоненциальное сглаживание и прогнозирование; X11(Census 2)-monthly-quarterly – X11 метод (месячно-квартально); Distributed lags analysis- анализ распределенных лагов; Spectral (Fourier) analysis – Спектральный (Фурье) анализ.

В разделе Replace missing data with представлены способы замены пропущенных данных: Overall mean – среднее значение выборки; Interpolation from adjacent points – интерполяция из смежных точек; Mean/Median of N adjacent points- среднее значение смежных точек; Predicted values from linear trend regression- предсказанное значение по линейному тренду.

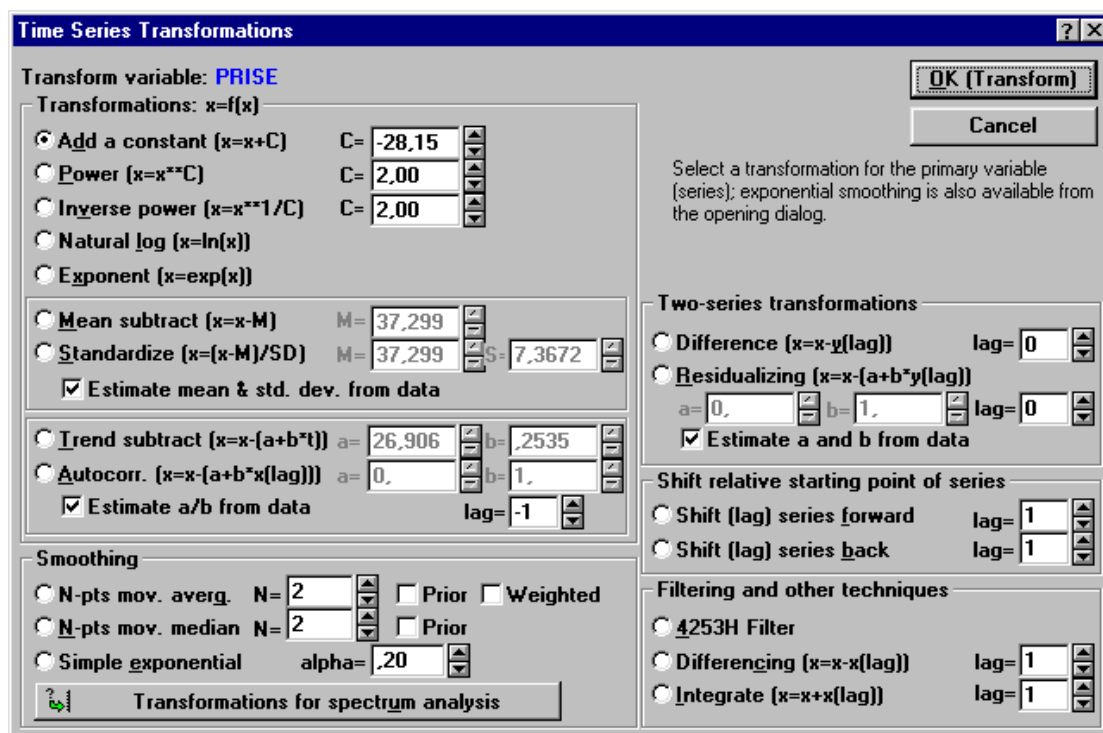


Рисунок 6.2. Диалоговое окно преобразования временного ряда.

Сглаживание временного ряда с помощью простой скользящей средней можно выполнить с помощью кнопки Ok (transformations, autocorrelations, crosscorrelations, plots) на стартовой панели. После нажатия кнопки появляется диалоговое окно Transformations of variables (трансформация переменных), в

котором надо указать переменную для трансформации. После нажатия на Ок появляется окно Time series transformations (преобразование временного ряда), в котором представлены разные способы преобразования переменных. В области Smoothing (сглаживание) можно задать сглаживание по простой нецентрированной (Prior) или взвешенной (Weighted) скользящей средней, простое экспоненциальное сглаживание (Simple exponential). В строке N-pts mov. averg/median указывают интервал сглаживания.

С помощью кнопки Open data (открыть данные) откроем файл с данными о цене открытия по акциям Газпрома на МФБ с 18.05.03 по 18.05.04 года (источник: [www.rbc.ru](http://www.rbc.ru)). Щелкнув по кнопке Variables (переменные), выберем переменную Prixe. Символ L возле имени переменной означает, что она закрыта на ключ, и переменную нельзя удалить. Кнопка Delete highlighted variable (удалить высвеченные переменные) позволяет удалить преобразованные (добавленные) переменные, но не исходные.

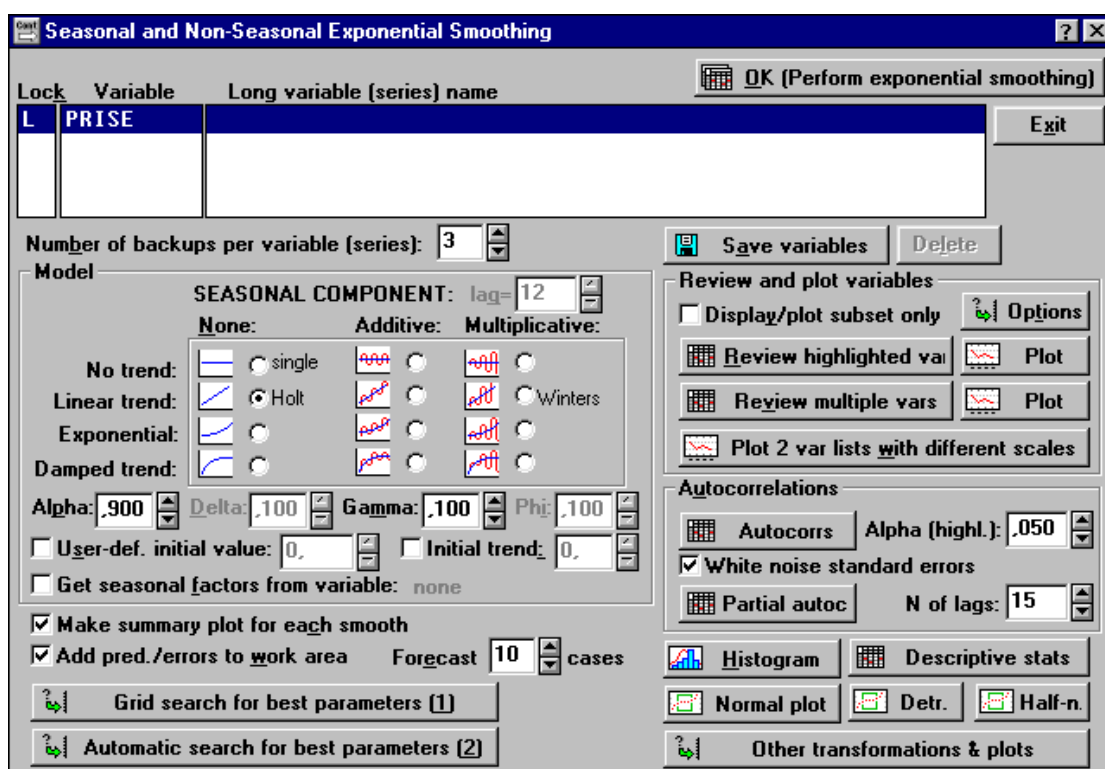


Рисунок 6.3. Стартовая панель Сезонное и несезонное сглаживание.

Для анализа данных выберем процедуру Exponential smoothing & forecasting – экспоненциальное сглаживание и прогнозирование. На экране появит-

ся стартовая панель Seasonal and Non-Seasonal Exponential Smoothing (сезонное и несезонное экспоненциальное сглаживание). Стартовая панель состоит из нескольких частей. В верхней части— область выбора переменной для анализа.

Ниже представлена область спецификации модели Model. Опишем ее подробнее. Чтобы выполнить экспоненциальное сглаживание без учета сезонных колебаний ряда, на панели предложены модели в столбце None. Для графической демонстрации результатов сглаживания установите флажок на кнопке Make summary plot for each Smooth (построить график результатов сглаживания). Если в таблице результатов требуется наличие предсказанных значений и остатков, то установите флажок на кнопке Add pred./errors to work area (добавить предсказанные значения и остатки в рабочую область). В строке Forecast (прогноз) задайте период прогнозирования.

В области Review and plot variables (обзор и графики переменных) можно просмотреть и изменить значения переменных (Review highlighted var), преобразовать переменные (Review multiple var), построить график (Plot).

В области Autocorrelations (автокорреляция) можно вывести автокорреляционную функцию временного ряда (Autocorr), и частную автокорреляционную функцию временного ряда (Partial auto).

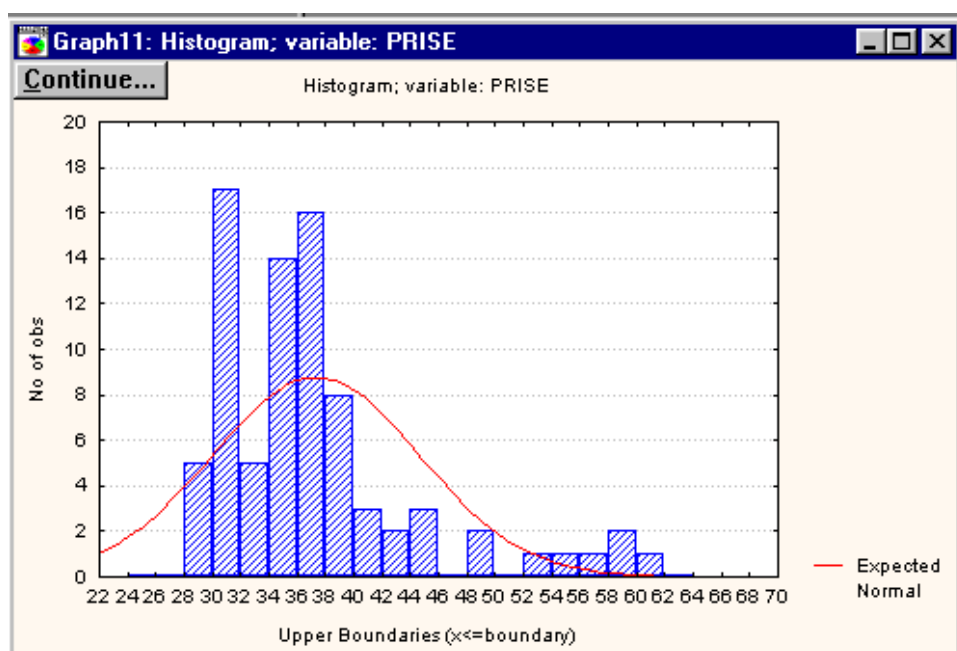


Рисунок 6.4. Гистограмма распределения цены открытия на акции «Газпром».

Очевидна асимметрия в выборке, наибольшее количество сделок заключалось по цене от 30 до 40 ден. единиц.

На стартовой панели также предложены кнопки для построения гистограммы, совмещенной в кривой нормального распределения; для расчета показателей описательной статистики; выполнения преобразований данных и построения других графиков. Для визуального определения типа тенденции во временном ряду построим его график. Щелкнем по верхней правой кнопке Plot на стартовой панели Seasonal and Non-Seasonal al Smoothing (Сезонное и несезонное сглаживание).

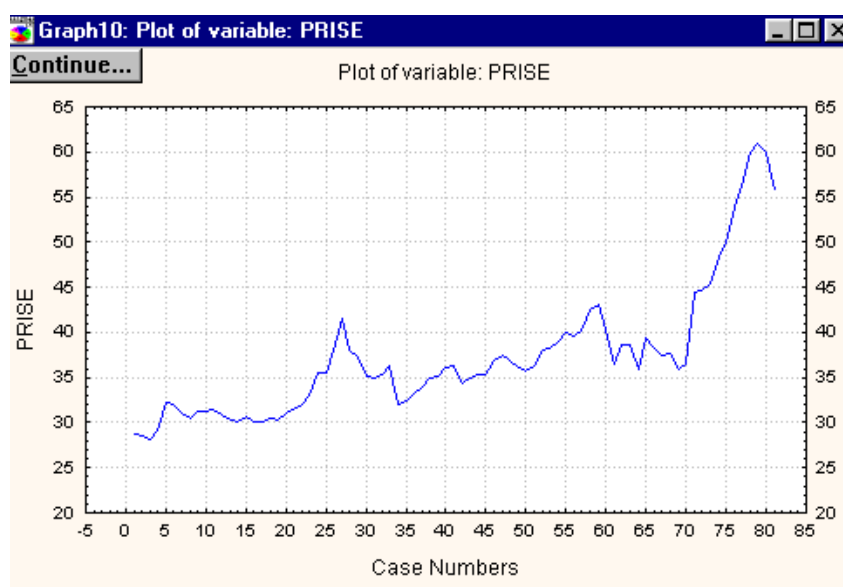


Рисунок 6.5. Динамика цены открытия на акции «Газпром» с 18.05.03 по 18.05.04.

На графике можно увидеть сезонные колебания с квартальной периодичностью и предположить наличие линейной тенденции.

Параметры экспоненциального сглаживания «альфа» и «гамма» по умолчанию равны 0,1. STATISTICA дает возможность автоматического поиска нужных параметров. Этому служит кнопка Grid search for best parameters (поиск по сетке лучших параметров). Щелкните на кнопку и на экране появится окно Parameter Grid Search (поиск параметров по сетке). В нем задаются начальные значения неизвестных параметров.

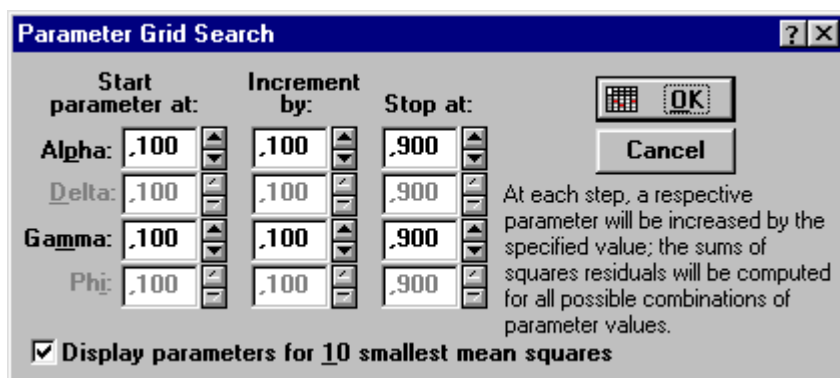


Рисунок 6.6. Окно поиска параметров по сетке.

В верхней строке даны лучшие значения: Alpha=0.9, Gamma=0.1.

Parameter grid search (Smallest abs. errors are highlighted)

Continue... Model: Linear trend, no season ; S0=28,63 T0=.3350  
PRISE

Model Number	Alpha	Gamma	Mean Error	Mean Abs Error	Sums of Squares	Mean Squares	Mean % Error	M %
73	.900000	.100000	.060454	1,274938	293,6986	3,625909	.044948	3
74	.900000	.200000	.006734	1,289480	303,7552	3,750064	-.014602	3
64	.800000	.100000	.079607	1,307436	305,4058	3,770442	.069501	3
75	.900000	.300000	-.027557	1,312587	312,0225	3,852129	-.061106	3
65	.800000	.200000	.018909	1,309953	313,5432	3,870904	.002680	3
66	.800000	.300000	-.021272	1,316015	319,3130	3,942136	-.052058	3
76	.900000	.400000	-.046925	1,359787	321,0187	3,963194	-.089823	3
55	.700000	.100000	.105860	1,367808	323,9090	3,998877	.103945	3
67	.800000	.400000	-.044689	1,341072	325,5084	4,018622	-.087481	3
56	.700000	.200000	.036879	1,351161	330,0715	4,074957	.029383	3

Рисунок 6.7. Таблица результатов поиска параметров по сетке.

Щелкнув на кнопку Continue (продолжить), вернитесь в окно Сезонное и несезонное экспоненциальное сглаживание и укажите лучшие значения параметров «альфа» и «гамма», Ок.

Exp. smoothing: S0=28,63 T0=.3350 (gasprom.sta)

Continue... Lin.trend, no season ; Alpha=.900 Gamma=.100  
PRISE

Case	PRISE	Smoothed Series	Resids
78	60,00000	57,46100	2,53900
79	61,00000	61,31991	-.31991
80	60,00000	62,57700	-2,57700
81	55,60000	61,57078	-5,97078
82		56,87278	
83			
84			
85			
86			
87			
88			
89			
90			
91			

TIME  
SERIES

Summary of error

Error
Mean error
Mean absolute error
Sums of squares
Mean square
Mean percentage error
Mean abs. perc. error

Lin.trend, no season ; Alpha=.900 Gamma=.100  
PRISE

.060454450154  
1,274938021522  
293,698595563004  
3,625908587198  
.044948153913  
3,289066011651

Рисунок 6.8. Таблицы результатов с прогнозной оценкой.



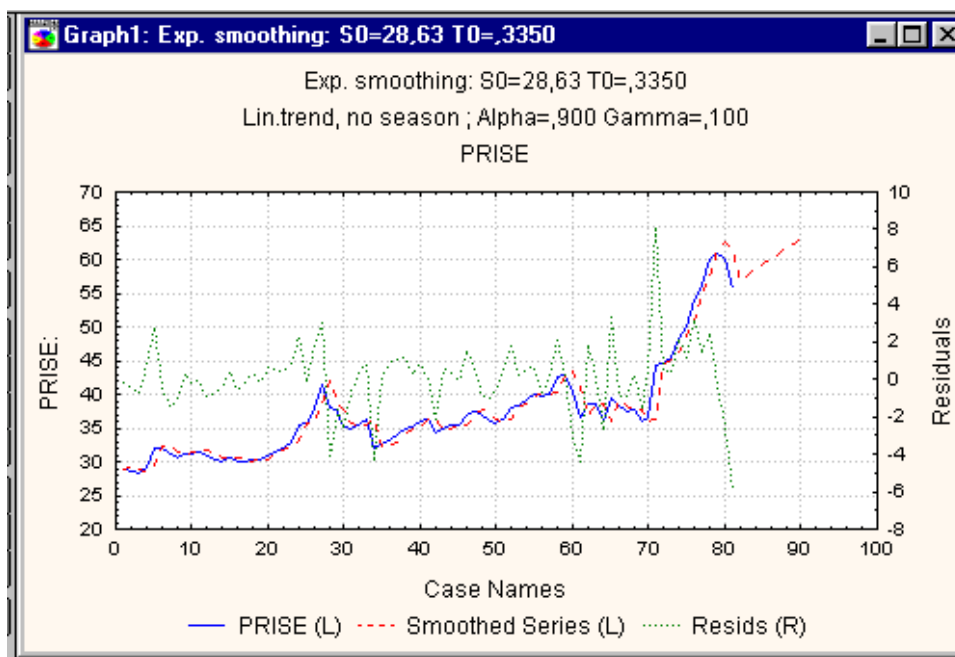


Рисунок 6.9. График наблюдаемых, сглаженных значений цены, прогнозной оценки и остатков.

Получим прогнозные оценки методом сезонной адаптивной экспоненциальной модели. На стартовой панели Seasonal and Non-Seasonal Exponential Smoothing (сезонное и несезонное экспоненциальное сглаживание) в области спецификации Model (модель) установите флажок на Additive (аддитивная) по строке Linear trend (линейный тренд). Выше было предположение о квартальной периодичности сезонных колебаний, поэтому в строке Seasonal Component (сезонная компонента) укажите лаг 4, Ок (Perform exponential smoothing).

Exp. smoothing: Additive season (4) S0=27,87 T0=,4032				
Continue... Lin.trend, add.season; Alpha=,901 Delta=,100 Gamma=,100				
Case	PRISE	Smoothed Series	Resids	Seasonal Factors
78	60,00000	57,44800	2,55200	
79	61,00000	61,81645	-,81645	
80	60,00000	62,26066	-2,26066	
81	55,60000	61,41277	-5,81277	
82		56,95992		
83		58,19500		
84		58,63310		
85		59,26296		
86		60,10444		
87		61,33952		
88		61,77762		
89		62,40748		
90		63,24896		
91		64,48404		

Рисунок 6.10. Результаты сглаживания с учетом сезонности.



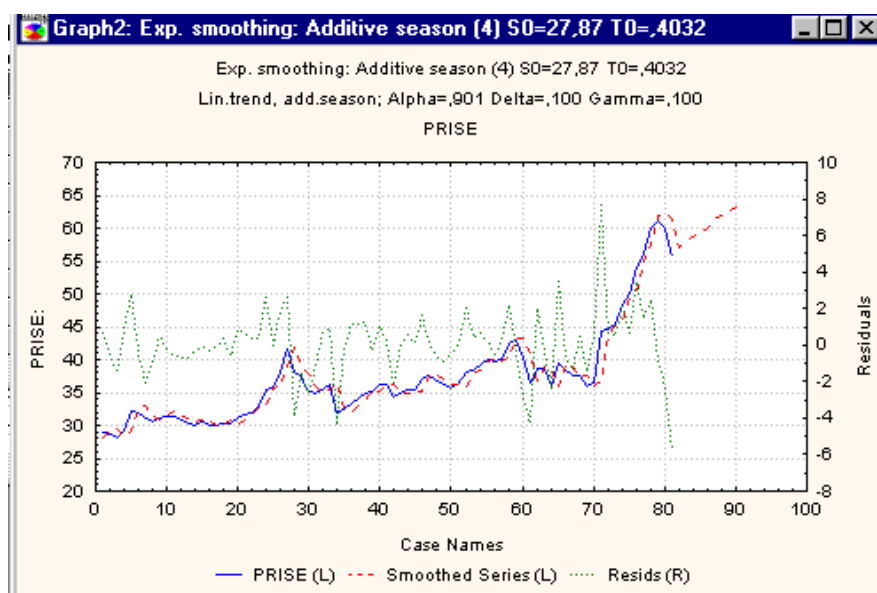


Рисунок 6.11. График наблюдаемых, сглаженных значений цены, прогнозной оценки и остатков с учетом сезонности.

Следует помнить, что экспоненциальное сглаживание наиболее простой метод прогнозирования. В данном методе не строятся доверительные интервалы и, следовательно, невозможно оценить риск при использовании прогноза. К этому методу следует обращаться на самом первом этапе исследования. Оценить подгонку модели поможет график остатков, который выводится вместе со сглаженным рядом, исходным рядом и прогнозом. В хорошо подогнанной модели в остатках не должно быть тенденции, зависимостей, увеличивающейся или уменьшающейся амплитуды колебаний.

#### Список литературы:

1. Макарова Н. В., Трофимец В. Я. Статистика в Excel: учебное пособие. – М.: Финансы и статистика, 2002 – 368 с.
2. Боровиков В. П. Программа Statistica для студентов и инженеров. – 2-е изд. – М.: КомпьютерПресс, 2001.-301 с. –ил.
3. Боровиков В. П., Ивченко Г. И., Прогнозирование в системе Statistica/ Учебное пособие – М.: Финансы и статистика, 1999. –384 с.: ил.
4. Боровиков В. П., Боровиков И. П. Statistica- статистический анализ и обработка данных в среде Windows. – М.: Филинь, 1998.-608 с.
5. Боровиков В. Statistica. Искусство анализа данных на компьютере: для профессионалов. 2-е изд. СПб.: Питер, 2003.- 688с.